UNIVERSITY OF Miami
Department of Electrical & Computer Engineering
College of Engineering
P.O. Box 248294
Coral Gables, Florida 33124-0640

TO

Final Technical Report

HIGH-SPEED FIXED- AND FLOATING-POINT IMPLEMENTATION OF DELTA-OPERATOR FORMULATED DISCRETE-TIME SYSTEMS

Principal Investigator: K. Premaratne
Grant No: N00014-94-1-0454
R&T Project Code: 3148508---01

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	24 February 1995	Final Technical; 01 Jan-31 Dec. 1994

4. TITLE AND SUBTITLE
High-speed fixed- and floating-point implementation of delta-operator formulated discrete-time systems

5. FUNDING NUMBERS
Grant No:
N00014-94-1-0454

R&T Project Code:
3148508---01

6. AUTHOR(S)
Kamal Premaratne

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Department of Electrical and Computer Engineering
University of Miami
1251 Memorial Drive, #EB406
Coral Gables, FL 33146
USA

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Office of Naval Research (ONR)
Code 251:JWK
Ballston Tower One
800 North Quincy Street
Arlington, VA 22217-5660 USA

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES
This project is a collaborative effort with Dr. Peter H. Bauer, Department of Electrical Engineering, University of Notre Dame, Notre Dame, IN 46556, USA, who is the principal investigator of Grant No: N00014-94-1-0387; R&T Project Code: 3148509---01.

12a. DISTRIBUTION/AVAILABILITY STATEMENT

DISTRIBUTION STATEMENT A
Approved for public release;
Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This final report addresses research results on three areas:
(a) Analysis and Design of Finite Wordlength Implementations of Linear, Time-Invariant Delta-Systems: With fixed-point, delta-operator based implementations are shown to always possess limit cycles regardless of quantization format. This problem is virtually non-existent with floating-point if mantissa is sufficiently long. When this is the case, delta-systems offer superior performance (especially, for sampled continuous-time systems) with high sampling rate.
(b) Analysis of Nonlinear Circuits Through Delta-Operator Based Schemes: Delta-operator based numerical schemes for nonlinear system simulation were proposed. For certain classes of nonlinearities, sensitivity measures and quantization error bounds for state orbit were also developed. With floating-point, the proposed numerical schemes are shown to produce superior performance when using a small discretization step size.
(c) 2-D and m-D Delta-Models: Delta-models for 2-D and m-D discrete-time systems were developed. Notions of gramians, balanced realizations, etc., were introduced. Coefficient sensitivity was also analyzed.
All results obtained were compared with corresponding results for the shift-operator case. Conditions where delta-systems are superior are also established.

14. SUBJECT TERMS
Delta-operator; discrete-time systems; finite wordlength effects; coefficient sensitivity; limit cycles; numerical simulation of nonlinear systems; 2-D and m-D digital filters.

15. NUMBER OF PAGES
151

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

GENERAL INSTRUCTIONS FOR COMPLETING SF 298

The Report Documentation Page (RDP) is used in announcing and cataloging reports. It is important that this information be consistent with the rest of the report, particularly the cover and title page. Instructions for filling in each block of the form follow. It is important to ***stay within the lines*** to meet ***optical scanning requirements***.

Block 1. Agency Use Only *(Leave blank)*.

Block 2. Report Date. Full publication date including day, month, and year, if available (e.g. 1 Jan 88). Must cite at least the year.

Block 3. Type of Report and Dates Covered. State whether report is interim, final, etc. If applicable, enter inclusive report dates (e.g. 10 Jun 87 - 30 Jun 88).

Block 4. Title and Subtitle. A title is taken from the part of the report that provides the most meaningful and complete information. When a report is prepared in more than one volume, repeat the primary title, add volume number, and include subtitle for the specific volume. On classified documents enter the title classification in parentheses.

Block 5. Funding Numbers. To include contract and grant numbers; may include program element number(s), project number(s), task number(s), and work unit number(s). Use the following labels:

C	- Contract	**PR**	- Project
G	- Grant	**TA**	- Task
PE	- Program Element	**WU**	- Work Unit Accession No.

Block 6. Author(s). Name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. If editor or compiler, this should follow the name(s).

Block 7. Performing Organization Name(s) and Address(es). Self-explanatory.

Block 8. Performing Organization Report Number. Enter the unique alphanumeric report number(s) assigned by the organization performing the report.

Block 9. Sponsoring/Monitoring Agency Name(s) and Address(es). Self-explanatory.

Block 10. Sponsoring/Monitoring Agency Report Number. *(If known)*

Block 11. Supplementary Notes. Enter information not included elsewhere such as: Prepared in cooperation with...; Trans. of...; To be published in.... When a report is revised, include a statement whether the new report supersedes or supplements the older report.

Block 12a. Distribution/Availability Statement. Denotes public availability or limitations. Cite any availability to the public. Enter additional limitations or special markings in all capitals (e.g. NOFORN, REL, ITAR).

- **DOD** - See DoDD 5230.24, "Distribution Statements on Technical Documents."
- **DOE** - See authorities.
- **NASA** - See Handbook NHB 2200.2.
- **NTIS** - Leave blank.

Block 12b. Distribution Code.

- **DOD** - Leave blank.
- **DOE** - Enter DOE distribution categories from the Standard Distribution for Unclassified Scientific and Technical Reports.
- **NASA** - Leave blank.
- **NTIS** - Leave blank.

Block 13. Abstract. Include a brief *(Maximum 200 words)* factual summary of the most significant information contained in the report.

Block 14. Subject Terms. Keywords or phrases identifying major subjects in the report.

Block 15. Number of Pages. Enter the total number of pages.

Block 16. Price Code. Enter appropriate price code *(NTIS only)*.

Blocks 17. - 19. Security Classifications. Self-explanatory. Enter U.S. Security Classification in accordance with U.S. Security Regulations (i.e., UNCLASSIFIED). If form contains classified information, stamp classification on the top and bottom of the page.

Block 20. Limitation of Abstract. This block must be completed to assign a limitation to the abstract. Enter either UL (unlimited) or SAR (same as report). An entry in this block is necessary if the abstract is to be limited. If blank, the abstract is assumed to be unlimited.

FINAL REPORT

to the

OFFICE OF NAVAL RESEARCH (ONR)

for support of research project

HIGH SPEED FIXED- AND FLOATING-POINT IMPLEMENTATION OF DELTA-OPERATOR FORMULATED DISCRETE-TIME SYSTEMS

Principal Investigator

Kamal Premaratne, PhD
Assistant Professor
Department of Electrical and Computer Engineering
University of Miami
1251 Memorial Drive, #EB406
Coral Gables, FL 33146 USA
Tel: +1.305.284.4051; .4044 (Fax)
Email: kprema@umiami.ir.miami.edu

Project Start Date: 01 January 1994
Project End Date: 31 December 1994

Project Duration: 12 months
Amount: $46,805

Grant Number: N00014-94-1-0454

R&T Project Code: 3148508—01

Contact at ONR: Dr. Clifford G. Lau, Tel: +1.703.696.4216
Contact at University of Miami: Ms. Sandy Blanco, Tel: +1.305.284.4541

Table of Contents

† Two-sided copies; but numbered only on one side.

I. Introduction

In many applications such as high speed digital signal processing, reliable simulations of dynamical systems, digital implementation and simulation of chaotic systems, etc., effects of finite wordlength are a critical issue. The process of actual digital computer implementation of a given ideal dynamical system can be characterized by several open parameters that have a critical impact on the performance of the actually implemented algorithm:

1. The realization (in the linear case, the system matrices): This determines the coefficients involved, the order of computation, etc. There are infinitely many realizations for implementing the same dynamical system.
2. The arithmetic format: This determines the type of arithmetic used (fixed-point, floating-point, etc.), the register lengths, and the type of quantization operations in the reformatting processes.

For the class of linear, time-invariant systems, δ-operator based implementations were shown to perform superior relative to their q-operator based counterparts, if the sampling rate is chosen sufficiently small [1,2]. These advantages of the δ-operator, especially in high speed, real-time applications, were demonstrated with respect to quantization noise at system output and differential sensitivity of the frequency response with respect to coefficients of system realization [1-4]. In addition, the use of δ-operators allows a unified treatment of both the continuous- and discrete-time cases. These properties make δ-operator based systems an attractive alternative to conventional system realizations.

However, a number of questions on δ-operator based implementations remained unanswered:

1. Do the advantages that have been demonstrated for the linear time-invariant case carry over to the 2-D, m-D, and possibly nonlinear cases? If so, δ-operator based numerical schemes can provide a completely novel, simple, widely applicable, yet more reliable methodology for system simulation and realization. The fundamental importance of such an investigation was identified at the very outset by the PI.
2. What about asymptotic stability of δ-operator systems and the possibility of limit cycles? Although quantization noise at the output was shown to be smaller for δ-system realizations, this does not automatically preclude the existence of limit cycles. In fixed-point implementations, the existence of prohibitively large limit cycles was evident. Although in almost all applications such behavior is unacceptable, no attention had been directed towards this seemingly generic phenomena of δ-systems.

The above questions are at the core of this research project. This report, which provides a description of the work carried out under this research project, is structured as follows: In Section II, a brief description of the proposed tasks is outlined. In Section III, the results obtained are briefly described on a qualitative level for

each of the problem areas tackled. Section IV offers conclusions and summarizes the accomplishments and their significance. Section V contains pertinent references.

More detailed technical decriptions of the results in Section III may be found in several technical papers/presentations, and these are included in Appendix A. It contains all those material that have already been published in or submitted to journals or conferences as well as those (such as, presentations, summaries, etc.) that has been submitted to ONR. Appendix B contains those technical papers/presentations that have some peripheral relevance to the proposed research, and those in which acknowledgement of ONR support is given.

II. Brief Description of Tasks

The proposed work was divided into three major tasks:

T1: Analysis and design of finite wordlength implementations of linear time-invariant δ-systems.

T2: Analysis of nonlinear circuits through δ-operator based schemes.

T3: 2-D and m-D δ-system models.

Task T1 reveals some fundamental difficulties in the implementation of δ-systems with fixed-point arithmetic. It focuses mainly on zero convergence of the free system response and exposes the existence of limit cycles as well as effects of sampling time Δ quantization.

Task T2 is a study of whether the superior finite wordlength properties associated with certain linear time-invariant system realizations also extend to nonlinear systems. This work was mainly motivated by some very promising simulation results of chaotic systems.

Task T3 develops the formalisms for 2-D and m-D system descriptions in δ-operator form. It also investigates sensitivity properties of these proposed m-D δ-models and compares them with conventional q-models.

III. Results and Accomplishments

This section offers brief qualitative descriptions of the results obtained during this project period. A more rigorous quantitative analysis of these results are to be found in Appendix A which contains all relevant technical papers/presentations.

III.1. Task T1: Analysis and Design of Finite Wordlength Implementations of Linear Time-Invariant δ-Systems

We have exposed a serious limitation of δ-operator based realizations of discrete-time systems: They cannot be free of limit-cycles when used with small sampling times and fixed-point arithmetic! In particular, DC limit cycles are always present when sampling time is smaller than 0.5 for rounding, and 1.0 for truncation. In other words, under these conditions, nonzero initial conditions can be found, such that the asymptotic response converges to an incorrect equilibrium point different from the origin [5]. This in fact is a generic problem with δ-systems in the sense that it is independent of the margin of stability (of the ideal linear system) and its realization. The main cause of this lies in the update equation where multiplication by sampling time (which is typically small) occurs. This results in a difference vector that quantizes to zero.

The use of novel quantization schemes with smaller deadzones was also shown to be ineffective: Although quantizers that significantly reduce DC limit cycle amplitude may be selected, new oscillatory limit cycles are usually created. A newly developed computer-aided search algorithm for the existence of limit cycles may be effectively used to investigate this phenomenon [6]. Through construction of deadband regions and simple bounding hypercubes, these limit cycle amplitudes have been shown to grow with increasing sampling rate [6,7]. Using results on necessary 1-D conditions for stability of m-D systems [8], m-D δ-systems were also shown to produce similar limit cycle behavior [7,9].

Another drawback of fixed-point δ-operator implementations is the required high dynamic range of coefficients and signals. This is due to the fact that, given a q-system, in obtaining the corresponding δ-system, a division by Δ (which is typically small) is involved. Hence, additional bits in coefficient/signal registers are generally required to avoid overflow [10].

The above investigations produced the following unavoidable conclusion: Since δ-operator formulated discrete-time systems are superior to their q-operator counterparts only when the sampling rate is chosen to be significantly smaller than one, fixed-point arithmetic is not a suitable format for δ-system implementations.

The situation is refreshingly different in floating-point arithmetic: The above mentioned problems (encountered under fixed-point arithmetic) vanish and δ-systems produce significant advantages under high speed conditions. We show that,

under floating-point format, a stable linear system (independent of realization) can always be implemented limit cycle free in the regular dynamic range [11]. Equivalently, limit cycles can always be restricted into underflow conditions. Such limit cycles are acceptable for most applications. Furthermore, the large dynamic range requirements of δ-systems may easily be accommodated in floating-point arithmetic.

For both fixed and floating-point systems, new differential sensitivity measures which are widely applicable even to nonlinear and time-variant systems were developed [10,12]. Instead of using sensitivity measures related to frequency response (as is the usual practice), a time domain approach using state-space methods was developed. Sensitivity of state trajectory with respect to system coefficients and initial conditions was investigated. For linear time-invariant systems, δ-operator based implementations have been shown to yield lower (by a factor Δ) sensitivity than their q-operator based counterparts. Sensitivities with respect to initial conditions are shown to be identical for both implementations.

III.2. Task T2: Analysis of Nonlinear Circuits Through δ-Operator Based Schemes

The following aspects of nonlinear δ-systems were addressed in detail:

(a) Sensitivity of state response with respect to coefficients of the nonlinear equation: This analysis was carried out for various types of nonlinearities as well as for both fixed and floating-point schemes [10,12].

(b) Bounds on quantization error magnitudes, required dynamic range, and construction of majorant systems for the response of δ-operator based implemented nonlinear systems [10].

In part (a), the concept of differential sensitivity of state response with respect to coefficients of the nonlinear equation was developed. The proposed sensitivity measures were evaluated for linear systems, piecewise linear systems, systems with $\mathcal{C}^1$ nonlinearities and systems with piecewise $\mathcal{C}^1$ nonlinearities.

For all these types of nonlinear systems, sensitivity of a δ-system with respect to coefficients was shown to be smaller (by a factor Δ) than that for its corresponding q-system under fixed-point arithmetic. For piecewise linear and piecewise $\mathcal{C}^1$ nonlinear systems, development of a quantitative measure for sensitivity of state trajectory with respect to initial conditions was required as well. This is due to how the piecewise characteristics of the nonlinearity is modelled. This proposed sensitivity measure was shown to be comparable for both q- and δ-systems.

Of course, nonlinear δ-systems, implemented in fixed-point arithmetic, can be shown to suffer from the same generic problem: Existence of incorrect equilibria. Since this is a serious problem especially in implementation and simulation of nonlinear systems, floating-point arithmetic was also intensively analyzed as an alternative. Suitable sensitivity measures were developed and evaluated for the non-

linear system types mentioned above. A comparison with corresponding q-operator based systems revealed what we believe to be a very important observation: Under mild conditions on the coefficients of the q-system, the state trajectory of the corresponding δ-system is less sensitive than that of the q-system. These conditions turn out to be routinely satisfied if the nonlinear discrete-time system is obtained through sampling of a given continuous-time system with a high sampling rate!

In part (b), a comparison between q- and δ-systems was conducted via quantization error bounds. For the fixed-point case, the q-system is always inferior to the δ-system, i.e., it produces larger quantization error bounds whenever single-length accumulators are used or when the number of computations in the state equation significantly exceeds the number of computations in the update equation.

Systems with polynomial type nonlinearities were investigated in great detail. For this class of nonlinearities, recommendations for the sampling rate which would provide an optimal balance between (a) the gains obtained from a reduced sampling period, and (b) the increased expense from a higher sampling frequency, are made. For sector bounded nonlinearities, majorant systems for the state response were constructed. When the sampling time is much smaller than 1, at each time instant, these majorant systems for δ-systems produce smaller state responses than those corresponding to q-systems.

For floating-point arithmetic, δ-systems produce smaller quantization error bounds than corresponding q-systems only when the nonlinearities satisfy certain magnitude conditions relative to the state vector. It was shown that, these conditions are always satisfied if the discrete-time system is produced by sampling the underlying continuous-time system at a very high rate. Note that, this is in accordance with our previous results on sensitivity. The underlying reason for these advantages of δ-systems is due to its implicit operand sorting. In other words, operands of similar 'size' are grouped together in the state equation of δ-systems, whereas, in the q-operator case, such a grouping is not implicit and a mix of operands of different 'sizes' is created.

III.3. Task T3: 2-D and m-D δ-System Models

In this task, the δ-operator counterpart to the 2-D Roesser q-model was developed [13]. It was shown that, for small sampling 'times' in both directions of propagation, the proposed 2-D and m-D models possess similar properties as the 1-D model. For example, fixed-point implementations are still plagued by limit cycles and not recommended; however, floating-point implementations can yield extremely attractive finite wordlength properties.

The usual system theoretic notions such as characteristic equation, transfer function, stability [14], etc., have been developed for the proposed 2-D δ-models. Furthermore, the notions of gramians, balanced realizations, and also its computa-

tion, were introduced.

To investigate coefficient sensitivity properties of 2-D δ-models, sensitivity measures appropriate for fixed and floating-point arithmetic schemes were developed. This analysis was carried out for the more general multi-input, multi-output case. The resulting conclusions may be summarized as follows:

1. In the fixed-point case, δ-models yield smaller coefficient sensitivity than the corresponding q-models when the sampling 'times' are small. Balanced realizations exhibit minimum coefficient sensitivity. This parallels the situation encountered in q-operator case. However, note that, generic limit cycle problems persist.
2. In the floating-point case, δ-models consistently offer superior coefficient sensitivity when the corresponding q-models' coefficients satisfy certain mild conditions. These conditions are routinely satisfied when implementing high-Q digital filters at high speeds. In most situations, 2-4 mantissa bits of an advantage is possible.

Furthermore, computation of balanced realizations has also been addressed. A simple relationship between balanced forms of corresponding q- and δ-system realizations has been established [13]. This makes it possible to derive balanced realizations using those algorithms that are applicable for the q-operator case.

IV. Conclusion

In summary, results obtained during the course of this funding period show that, δ-operator implementations of discrete-time systems can be quite superior to their q-operator counterparts if they are used correctly. We have shown that, great gains can be achieved in the case when a continuous-time system is sampled at a very high rate and is implemented in floating-point arithmetic. Similar comments are applicable to nonlinear and m-D systems as well.

Based on this work, we may make the following conclusion: δ-operator based implementations offer a number of unique and desirable properties which are essential in high performance applications, such as, high speed DSP and reliable simulations of dynamical systems. For such applications (where traditional q-operator based implementations are known to be ill-conditioned), δ-operator based schemes provide a general and easily applicable technique for reliable implementation of discrete-time systems.

V. References

[1] Goodwin, G.C., Middleton, R.H., and Poor, H.V., "High speed digital signal processing and control," *Proc. IEEE,* **80**, pp. 240-259, 1992.

[2] Middleton, R.H., and Goodwin, G.C., *Digital Control and Estimation: A Unified Approach,* Englewodd Cliffs, NJ: Prentice Hall, 1988.

[3] Li, G., and Gevers, M., "Comparative study of finite wordlength effects in shift and delta operator parameterization," *Proc. IEEE CDC'90,*, **2**, pp. 954-959, Honolulu, HI, 1990.

[4] Li, G., and Gevers, M., "Roundoff noise minimization using delta-operator realizations," *IEEE Trans. Sig. Proc.,* **41**, pp. 629-637, 1993.

[5] Premaratne, K., and Bauer, P.H., "Limit cycles and asymptotic stability of delta-operator systems in fixed-point arithmetic," *Proc. IEEE ISCAS'94,* **2**, pp. 461-464, London, UK, 1994.

[6] Premaratne, K., Kulasekere, E.C., Bauer, P.H., and Leclerc, L.J., "An exhaustive search algorithm for checking limit cycle behavior of digital filters," *IEEE Trans. Sig. Proc.,* 1995, in review.

[7] Bauer, P.H., and Premaratne, K., "Limit cycles in delta-operator formulated 1-D and m-D discrete-time systems with fixed-point arithmetic," *IEEE Trans. Circ. Syst.—I: Fund. Theo. Appl.,* 1995, in review.

[8] Bauer, P.H., "Low-dimensional conditions for global asymptotic stability of m-D nonlinear digital filters," *Proc. IEEE ISCAS'94,* **2**, pp. 553-556, London, UK, 1994.

[9] Bauer, P.H., and Premaratne, K., "Fixed-point implementations of m-D delta-operator formulated discrete-time systems: Difficulties in convergence," *Proc. IEEE SOUTHEASTCON'94,* pp. 26-29, Miami, FL, 1994.

[10] *Progress Reports I and II,* presentations to Dr. C.G. Lau, ONR, Arlington, VA, Sept. 1994.

[11] Bauer, P.H., and Premaratne, K., "Zero-convergence of 2-D Roesser state-space models implemented in floating-point arithmetic," *Proc. 38th Midwest Symp. on Circ. and Syst. (MWSCAC'95),* Rio de Janeiro, Brazil, 1995, to appear.

[12] Premaratne, K., and Bauer, P.H., "Digital simulation of nonlinear systems using delta-operator based numerical schemes," *IASTED Int. Conf. Modelling and Simulation,* Colombo, Sri Lanka, 1995, in review.

[13] Premaratne, K., Ekanayake, M.M., Suarez, J., and Bauer, P.H., "Two-dimensional delta-operator formulated discrete-time systems: State-space realization and its coefficient sensitivity properties," *IEEE Trans. Sig. Proc.* 1995, in review.

[14] Premaratne, K., and Boujarwah, A.S., "An algorithm for stability determination of two-dimensional delta-operator formulated discrete-time systems," *Multidim. Syst. Sig. Proc.,* 1995, to appear.

Appendix A: Papers/Presentations Directly Related to Grant #N00014-94-1-0454

List of Papers/Presentations

Journal papers:

[A1] Premaratne, K., and Boujarwah, A.S., "An algorithm for stability determination of two-dimensional delta-operator formulated discrete-time systems," *Multidim. Syst. Sig. Proc.,* 1995, to appear.

[A2] Premaratne, K., Kulasekere, E.C., Bauer, P.H., and Leclerc, L.J., "An exhaustive search algorithm for checking limit cycle behavior of digital filters," *IEEE Trans. Sig. Proc.,* 1995, in review.

[A3] Bauer, P.H., and Premaratne, K., "Limit cycles in delta-operator formulated 1-D and m-D discrete-time systems with fixed-point arithmetic," *IEEE Trans. Circ. Syst.—I: Fund. Theo. Appl.,* 1995, in review.

[A4] Premaratne, K., Ekanayake, M.M., Suarez, J., and Bauer, P.H., "Two-dimensional delta-operator formulated discrete-time systems: State-space realization and its coefficient sensitivity properties," *IEEE Trans. Sig. Proc.* 1995, in review.

Conference papers:

[A5] Bauer, P.H., and Premaratne, K., "Fixed-point implementations of m-D delta-operator formulated discrete-time systems: Difficulties in convergence," *Proc. IEEE SOUTHEASTCON'94,* pp. 26-29, Miami, FL, Apr. 1994.

[A6] Premaratne, K., and Bauer, P.H., "Limit cycles and asymptotic stability of delta-operator systems in fixed-point arithmetic," *Proc. IEEE ISCAS'94,* Vol. 2, pp. 461-464, London, UK, May-June 1994.

[A7] Premaratne, K., Ekanayake, M.M., Suarez, J., and Bauer, P.H., "Two-dimensional delta-operator formulated discrete-time systems: State-space realization and its coefficient sensitivity properties," *Proc. 37th Midwest Symp. Circ. Syst. (MWSCAS'94),* pp. 805-808, Lafayette, LA, Aug. 1994.

[A8] Premaratne, K., Ekanayake, M.M., and Bauer, P.H., "On balanced realizations of 2-D delta-operator formulated discrete-time systems," *Proc. SOUTHCON'95,* Fort Lauderdale, FL, 1995, to appear.

[A9] Premaratne, K., Kulasekere, E.C., Bauer, P.H., and Leclerc, L.J., "An exhaustive search algorithm for checking limit cycle behavior of digital filters," *Proc. IEEE ISCAS'95,* Seattle, WA, April-May 1995, to appear.

[A10] Bauer, P.H., and Premaratne, K., "Zero-convergence of 2-D Roesser state-space models implemented in floating-point arithmetic," *Proc. 38th Midwest Symp. on Circ. and Syst. (MWSCAC'95),* Rio de Janeiro, Brazil, Aug. 1995, to appear.

[A11] Premaratne, K., and Bauer, P.H., "Digital simulation of nonlinear systems using delta-operator based numerical schemes," *Proc. IASTED Int. Conf. Modelling and Simulation,* Colombo, Sri Lanka, July 1995, in review.

Presentations:

[A12] *Semiannual Performance Report,* Aug. 1994.

[A13] *Progress Report I,* presentation to Dr. C.G. Lau, ONR, Arlington, VA, Sept. 1994.

[A14] *Progress Report II,* presentation to Dr. C.G. Lau, ONR, Arlington, VA, Sept. 1994.

[A15] *Progress Report/Viewgraphs,* prepared for Dr. C.G. Lau, ONR, Arlington, VA, Nov. 1994.

To appear in Multidimensional Systems and Signal Processing, 1995

jun0193*.tex

An Algorithm for Stability Determination of Two-Dimensional Delta-Operator Formulated Discrete-Time Systems

KAMAL PREMARATNE
Department of Electrical and Computer Engineering, University of Miami, P.O. Box 248294, Coral Gables, FL 33124, U.S.A.

A.S. BOUJARWAH
Electrical and Computer Engineering Department, College of Engineering and Petroleum, Kuwait University, P.O. Box 5969, 13060 Safat, Kuwait.

Abstract. The recent interest in delta-operator (or, δ-operator) formulated discrete-time systems (or, δ-systems) is due mainly to (a) their superior finite wordlength characteristics as compared to their more conventional shift-operator (or, q-operator) counterparts (or, q-systems), and (b) the possibility of a more unified treatment of both continuous- and discrete-time systems. With such advantages, design, analysis, and implementation of two-dimensional (2-D) discrete-time systems using the δ-operator is indeed warranted. Towards this end, the work in this paper addresses the development of an easily implementable *direct* algorithm for stability checking of 2-D δ-system transfer function models. *Indirect* methods that utilize transformation techniques are not pursued since they can be numerically unreliable. In developing such an algorithm, a tabular form for stability checking of δ-system characteristic polynomials with complex-valued coefficients and certain quantities that may be regarded as their corresponding Schur-Cohn minors are also proposed.

Keywords. Two-dimensional discrete-time systems, two-dimensional digital filters, δ-operator formulated discrete-time systems, bivariate polynomials, Schur-Cohn minors, stability.

1. Introduction

The increased interest in δ-systems during the recent years (see [1-6], and references therein) is due mainly to two reasons: (a) δ-systems provide superior finite wordlength properties with respect to roundoff noise propagation [5] and coefficient sensitivity [1], [5], [7], as compared to their q-system counterparts, and (b) the δ-operator yields the differential operator as a limiting case when sampling time approaches zero enabling a unified treatment of both continuous- and discrete-time systems [1].

With such advantages in mind, development of 2-D and multi-dimensional (m-D) δ-system models must clearly be undertaken. Such research can, for example, provide m-D digital filters with superior roundoff error and coefficient sensitivity performance allowing their implementation to be carried out in a shorter wordlength environment. This is especially crucial in real-time applications, such as, in implementing narrow bandwidth filters under high sampling rates (for example, in current wide bandwidth communication system applications) where traditional q-operator implementations perform poorly [8].

In applications mentioned above, and those dealing with high-speed processing of 2-D and m-D data (for instance, in weather, seismic, gravitational photographs, video images, systems with mutliple sampling rates, etc.), ensuring stability is an important consideration (see [9], and references therein). Given the characteristic polynomial of a δ-system, to determine stability, one may first use a variable transformation that yields a more familiar stability region, for instance, the unit bi-circle. Then, an existing technique (see [9-10], and references therein) may be applied. However, such techniques are known to be prone to numerically ill-conditioning [1], [6]. In the 1-D case, direct stability checking methods for δ-system polynomials are in [6] (where a tabular method based on the work in [11] is given) and [12] (where a Hermite-Bieler-like Theorem is utilized). Hence, our purpose here is to develop a *direct* easily implementable stability checking technique applicable to m-D δ-systems. As usual, for notational simplicity, we concentrate on the 2-D case, the extension to the m-D case being quite straight-forward.

In checking stability of bivariate characteristic polynomials, two conditions must be

satisfied.

(a) Condition I involves a 1-D stability check of a polynomial with real-valued coefficients. One may use the table form in [6]. Alternately, one may utilize an explicit root location scheme.

(b) Condition II involves a stability check of a polynomial with complex-valued coefficients where the latter are dependent on a parameter taking values on a certain circle in the complex plane. Explicit root location schemes are now ineffective, and the value of tabular methods becomes apparent. Note that, in such a situation, compared to Nyquist-like techniques [13], tabular methods are known to provide certain numerical advantages as well [14].

In checking condition II for 2-D q-systems, an effective technique involves checking positive definiteness of the Hermitian Schur-Cohn matrix [15]. This lets one use an important simplification due to Siljak [16]. The tabular form in [15] takes full use of this since it provides the Schur-Cohn minors (that is, the principal minors of the Hermitian Schur-Cohn matrix) directly from its entries [15], [17]. A similar simplification applicable to δ-systems is clearly possible if condition II may be reduced to checking positive definiteness of a Hermitian matrix.

With the above in mind, we develop the following in this paper: (a) Tabular form for stability checking of δ-system characteristic polynomials possessing complex-valued coefficients, (b) Analogs of Schur-Cohn minors and a corresponding Hermitian matrix applicable for such systems, and (c) a direct stability checking algorithm for 2-D δ-system transfer function models.

The paper is organized as follows. Section 2 introduces the notation used throughout and a brief review of previous results. Section 3 develops a tabular form for stability checking of δ-systems with complex-valued coefficients and some important relevant results. Section 4 presents quantities that may be regarded as the analogs of Schur-Cohn minors for δ-systems. The 2-D stability checking algorithm in Section 5 is based on the tabular form for real-valued coefficients [6]. Since only little extra work is needed, results in both

Sections 3 and 4 however are developed for the more general complex-coefficient case. Section 6 presents an example to validate the results. Section 7 contains the conclusion and some final remarks.

2. Preliminaries

2.1. Notation

$\Re$, $\Im$	Real and complex number fields.
$\Re^{p\times q}$, $\Im^{p\times q}$	Set of matrices of size $p\times q$ over $\Re$ and $\Im$, respectively.
$\mathrm{var}\{\cdot\}$	Number of sign changes in the sequence $\{\cdot\}$ of real numbers.
$\mathrm{Re}[\cdot]$, $\mathrm{Im}[\cdot]$	Real part and imaginary part of $[\cdot]\in\Im$.
$\bar{[\cdot]}$	Complex conjugate of $[\cdot]\in\Im$.
A^T, $\bar{A}$, A^*	Transpose, complex conjugate, and complex conjugate transpose of $A\in\Im^{p\times q}$, respectively.
$\Re[w]_n$, $\Im[w]_n$	Set of univariate polynomials of degree n (with respect to the indeterminate $w\in\Im$) over $\Re$ and $\Im$, respectively.
$\Re(w)$	Set of rational univariate polynomials (that is, quotient of univariate polynomials) over $\Re$.
$\Re[w_1]_{n_1}[w_2]_{n_2}$	Set of bivariate polynomials of relative degrees n_1 and n_2 (with respect to the indeterminates $w_1\in\Im$ and $w_2\in\Im$, respectively) over $\Re$.
$\Re(w_1,w_2)$	Set of rational bivariate polynomials over $\Re$.
z, c	Indeterminates of q- and δ-systems, respectively.
τ	Real positive number, usually the sampling time.

The transformation relationship between corresponding q-and δ-systems is

$$\delta=\frac{q-1}{\tau}\Longleftrightarrow c=\frac{z-1}{\tau}. \tag{2.1}$$

$\breve{[\cdot]}$	q-system quantity analogous to its corresponding δ-system quantity $[\cdot]$; for example, transfer function of a given discrete-time system is either $H(c)$ if implemented based on the δ-operator or $\breve{H}(z)$ if implemented based on the q-operator.
$H(c)\vert_{c\to z}$	$H(c)\vert_{c=(z-1)/\tau}$
$G(z)\vert_{z\to c}$	$G(z)\vert_{z=1+\tau c}$
$H(c_1,c_2)\vert_{\mathbf{c}\to\mathbf{z}}$	$H(c_1,c_2)\vert_{c_i=(z_i-1)/\tau,\,i=1,2}$
$G(z_1,z_2)\vert_{\mathbf{z}\to\mathbf{c}}$	$G(z_1,z_2)\vert_{z_i=1+\tau c_i,\,i=1,2}$

Stability studies of 1-D and 2-D q- and δ-systems involve the following regions:

$\mathcal{U}_q, \mathcal{U}_q^2$ $\{z \in \Im : |z| < 1\}$, $\{(z_1, z_2) \in \Im^2 : |z_i| < 1,\ i = 1, 2\}$.

$\overline{\mathcal{U}}_q, \overline{\mathcal{U}}_q^2$ $\{z \in \Im : |z| \leq 1\}$, $\{(z_1, z_2) \in \Im^2 : |z_i| \leq 1,\ i = 1, 2\}$.

$\mathcal{T}_q, \mathcal{T}_q^2$ $\{z \in \Im : |z| = 1\}$, $\{(z_1, z_2) \in \Im^2 : |z_i| = 1,\ i = 1, 2\}$.

$\mathcal{U}_\delta, \mathcal{U}_\delta^2$ $\{c \in \Im : |c + 1/\tau| < 1/\tau\}$, $\{(c_1, c_2) \in \Im^2 : |c_i + 1/\tau| < 1/\tau,\ i = 1, 2\}$.

$\overline{\mathcal{U}}_\delta, \overline{\mathcal{U}}_\delta^2$ $\{c \in \Im : |c + 1/\tau| \leq 1/\tau\}$, $\{(c_1, c_2) \in \Im^2 : |c_i + 1/\tau| \leq 1/\tau,\ i = 1, 2\}$.

$\mathcal{T}_\delta, \mathcal{T}_\delta^2$ $\{c \in \Im : |c + 1/\tau| = 1/\tau\}$, $\{(c_1, c_2) \in \Im^2 : |c_i + 1/\tau| = 1/\tau,\ i = 1, 2\}$.

To avoid unnecessary notational complications, the sampling time in both horizontal and vertical directions is taken to be equal to $\tau > 0$.

To emphasize the degree of $F(w) = \sum_{k=0}^{n} a_k^{(n)} w^k \in \Im[w]_n$, we sometimes denote it as $F(w)_n$ as well.

$\bar{F}(w)$ Conjugate polynomial of $F(w)$, that is, $\sum_{k=0}^{n} \bar{a}_k^{(n)} w^k$

$F^\sharp(z)$ Reciprocal polynomial of $F(z)$, that is, $z^n \bar{F}(1/z)$

$F^\sharp(c)$ Reciprocal polynomial of $F(c)$, that is, $(1 + \tau c)^n \bar{F}\left(\frac{-c}{1+\tau c}\right)$

A q-system polynomial is *q-symmetric* if $F(z) = F^\sharp(z)$. A δ-system polynomial is *δ-symmetric* if $F(c) = F^\sharp(c)$.

Tabular forms of stability checking of a polynomial in $\Im[\omega]_n$ typically employ a sequence of polynomials each of descending order. The first row of such a tabular form is denoted as *row* $\#n$, the second row is *row* $\#n - 1$, and so on.

JT, MJT Jury table [18], modified Jury table [15], [17].

real-q-BT Bistritz table for q-system polynomials with real-valued coefficients [11].

complex-q-BT Bistritz table for q-system polynomials with complex-valued coefficients [19].

real-δ-BT Table form for δ-system polynomials with real-valued coefficients [6].

complex-δ-BT Table form for δ-system polynomials with complex-valued coefficients (this paper).

A q-system polynomial with all its roots in $\mathcal{U}_q$ (for the 1-D case) or $\mathcal{U}_q^2$ (for the 2-D case) is said to be *stable*. The corresponding regions for a δ-system polynomial are $\mathcal{U}_\delta$ (for the 1-D case) or $\mathcal{U}_\delta^2$ (for the 2-D case), respectively.

2.2. Review of complex-q-BT

The complex-δ-BT introduced in Section 3 is based on the complex-q-BT, and hence, we briefly review it now. For more details, see [10]. Let the characteristic polynomial of a q-system be

$$\breve{F}(z) = \sum_{k=0}^{n} \breve{a}_k^{(n)} z^k \in \Im[z]_n \quad \text{with} \quad \breve{F}(1) \in \Re \quad \text{and} \quad \breve{F}(1) \neq 0. \tag{2.2}$$

The complex-q-BT is formed using the symmetric polynomial sequence $\{\breve{T}(z)_i\}_{i=0}^n$ where [19]

$$\breve{T}(z)_i = \begin{cases} \breve{F}(z)_n + \breve{F}^\sharp(z)_n, & \text{for } i = n; \\ \dfrac{\breve{F}(z)_n - \breve{F}^\sharp(z)_n}{z-1}, & \text{for } i = n-1; \\ \dfrac{(\breve{\delta}_{i+2} + \bar{\breve{\delta}}_{i+2} z)T(z)_{i+1} - T(z)_{i+2}}{z}, & \text{for } i = n-2, n-3, \ldots, 0, \end{cases} \tag{2.3}$$

where

$$\breve{\delta}_{i+2} = \frac{\breve{T}(0)_{i+2}}{\breve{T}(0)_{i+1}} = \frac{\breve{t}_0^{(i+2)}}{\breve{t}_0^{(i+1)}}, \; i = n-2, n-3, \ldots, 0. \tag{2.4}$$

As in [11] and [19], equating similar powers on either side, we may also get the following determinental rule: For $k = 0, 1, \ldots, i$, and $i = n-2, n-3, \ldots, 0$,

$$\breve{t}_k^{(i)} = \frac{1}{\breve{t}_0^{(i+1)}} \begin{vmatrix} \breve{t}_0^{(i+2)} & \breve{t}_{k+1}^{(i+2)} \\ \breve{t}_0^{(i+1)} & \breve{t}_{k+1}^{(i+1)} \end{vmatrix} + \frac{1}{\breve{t}_{i+1}^{(i+1)}} \begin{vmatrix} \breve{t}_{i+2}^{(i+2)} & \breve{t}_{k+1}^{(i+2)} \\ \breve{t}_{i+1}^{(i+1)} & \breve{t}_k^{(i+1)} \end{vmatrix} + \breve{t}_{k+1}^{(i+2)}. \tag{2.5}$$

Remark. The computational advantage of BT is due to $\breve{T}(z)_i$ being q-symmetric. This implies $\breve{t}_k^{(i)} = \bar{\breve{t}}_{i-k}^{(i)}$, $k = 0, 1, \ldots, i$, and hence, it is necessary to evaluate only half the coefficients of each row.

Using (12-13), (16), and Theorem 6 of [19], we get

THEOREM 2.1. [19] The polynomial $\breve{F}(z) \in \Im[z]_n$ is q-stable iff
I. $\breve{t}_0^{(i)} \neq 0$, $i = n-1, n-2, \ldots, 0$, and
II. $\nu_n = \text{var}\{\breve{T}(1)_n, \breve{T}(1)_{n-1}, \ldots, \breve{T}(1)_0\} = 0$.

2.3. Some results on 2-D stability

Consider the 2-D q-system transfer function

$$\breve{H}(z_1, z_2) = \frac{\breve{E}(z_1, z_2)}{\breve{F}(z_1, z_2)} \in \Re(z_1, z_2) \tag{2.6}$$

where $\breve{E}(z_1, z_2) \in \Re[z_1]_{n_1}[z_2]_{n_2}$ and $\breve{F}(z_1, z_2) \in \Re[z_1]_{n_1}[z_2]_{n_2}$. The 2-D z-transform is taken using positive powers of z_i. For a comprehensive discussion regarding stability of such systems, see [9-10], and references therein. Hence, for reasons of brevity, only some analog results applicable to 2-D δ-systems are provided. It is only necessary to observe that the corresponding δ-system $H(c_1, c_2)$ satisfies

$$H(c_1, c_2) = \frac{E(c_1, c_2)}{F(c_1, c_2)} = \breve{H}(z_1, z_2)|_{\mathbf{z}\to\mathbf{c}} \in \Re(c_1, c_2) \tag{2.7}$$

where $E(c_1, c_2) \in \Re[c_1]_{n_1}[c_2]_{n_2}$ and $F(c_1, c_2) \in \Re[c_1]_{n_1}[c_2]_{n_2}$. In the remainder of this paper, we will only be dealing with transfer functions $H(c_1, c_2)$ that are devoid of nonessential singularities of the second kind on $\mathcal{T}_\delta^2$ and the pair $E(c_1, c_2)$ and $F(c_1, c_2)$ is taken to be coprime. If the 2-D polynomial $F(c_1, c_2) \neq 0$, $\forall (c_1, c_2) \in \overline{\mathcal{U}}_\delta^2$, it is said to be *$\delta$-stable*. After using (2.1), the following result follows directly from [20]:

THEOREM 2.2. The 2-D δ-system in (2.7) is δ-stable iff
I. $F(c_1, -1/\tau) \neq 0$, $\forall c_1 \in \overline{\mathcal{U}}_\delta$, and
II. $F(c_1, c_2) \neq 0$, $\forall c_1 \in \mathcal{T}_\delta$, $\forall c_2 \in \overline{\mathcal{U}}_\delta$.

The following result, which allows one to use the real-δ-BT, is directly from [21-22] after using (2.1):

THEOREM 2.3. The 2-D δ-system in (2.7) is δ-stable iff
I. $F(c_1, -1/\tau) \neq 0$, $\forall c_1 \in \overline{\mathcal{U}}_\delta$, and

II. $G(x, c_2) \neq 0$, $\forall x \in [-2/\tau, 0]$, $\forall c_2 \in \overline{\mathcal{U}}_\delta$.
Here $G(x, c_2) = F(c_1, c_2)F(\bar{c}_1, c_2)\Big|_{\substack{c_1 \in \mathcal{T}_\delta \\ x=(c_1+\bar{c}_1)/2}}$.

2.4. Schur-Cohn minors

In stability checking of 2-D q-systems, the following result is important:

THEOREM 2.4. [15], [23-24] The polynomial $\breve{F}(z) \in \Im[z]_n$ is stable iff $\breve{\Delta}_i > 0$, $i = 1, 2, \ldots, n$, where $\breve{\Delta}_i$ is the principal minor of the Hermitian Schur-Cohn matrix $\breve{\Gamma} = \breve{\Gamma}^* = \{\breve{\gamma}_{ij}\} \in \Im^{n\times n}$ defined as

$$\breve{\gamma}_{ij} = \sum_{k=1}^{i} (\breve{a}_{n-i+k}\bar{\breve{a}}_{n-j+k} - \bar{\breve{a}}_{i-k}\breve{a}_{j-k}), \quad \text{for} \quad i \leq j.$$

Stability checking of 2-D q-systems then involves positivity checking of all Schur-Cohn minors $\breve{\Delta}_i(z)$, $\forall i = 1, 2, \ldots, n$, $\forall |z| = 1$. A necessary and sufficient condition for this is positivity of $\breve{\Delta}_i(1)$, $\forall i = 1, 2, \ldots, n$, and $\breve{\Delta}_n(z)$, $\forall |z| = 1$. This is the simplification due to [16] that has been effectively utilized in applying the MJT [15]. The advantage of the latter is that its entries yield the Schur-Cohn minors directly. The fact that complex-q-BT's entries also yield the Schur-Cohn minors was only recently shown.

THEOREM 2.5. [10], [25] The Schur-Cohn minors of $\breve{F}(z)$ are the principal minors of the $(n \times n)$ tridiagonal Hermitian matrix

$$\breve{\Delta} = \begin{bmatrix} \mathrm{Re}[\breve{t}_0^{(n)}\breve{t}_{n-1}^{(n-1)}] & \frac{1}{2}[\breve{t}_{n-1}^{(n-1)}\breve{t}_0^{(n-2)}] & 0 & \cdots & 0 & 0 \\ \frac{1}{2}[\breve{t}_0^{(n-1)}\breve{t}_{n-2}^{(n-2)}] & \mathrm{Re}[\breve{t}_0^{(n-1)}\breve{t}_{n-2}^{(n-2)}] & \frac{1}{2}[\breve{t}_{n-2}^{(n-2)}\breve{t}_0^{(n-3)}] & \ddots & 0 & 0 \\ 0 & \frac{1}{2}[\breve{t}_0^{(n-2)}\breve{t}_{n-3}^{(n-3)}] & \mathrm{Re}[\breve{t}_0^{(n-2)}\breve{t}_{n-3}^{(n-3)}] & \ddots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \ddots & \mathrm{Re}[\breve{t}_0^{(2)}\breve{t}_1^{(1)}] & \frac{1}{2}[\breve{t}_1^{(1)}\breve{t}_0^{(0)}] \\ 0 & 0 & 0 & \cdots & \frac{1}{2}[\breve{t}_0^{(1)}\breve{t}_0^{(0)}] & \mathrm{Re}[\breve{t}_0^{(1)}\breve{t}_0^{(0)}] \end{bmatrix}.$$

3. Complex-δ-BT

With no loss of generality, consider the δ-system characteristic polynomial

$$F(c) = \sum_{k=0}^{n} a_k^{(n)} c^k \in \Im[c]_n, \tag{3.1}$$

where

$$a_0^{(n)} \in \Re \quad \text{and} \quad a_0^{(n)} > 0. \tag{3.2}$$

We now construct the complex-δ-BT with the use of the δ-symmetric polynomial sequence $\{T(c)_i\}_{i=0}^{n}$ where

$$T(c)_i = \begin{cases} F(c)_n + F^{\sharp}(c)_n, \ i = n; \\ \dfrac{F(c)_n - F^{\sharp}(c)_n}{c}, \ i = n-1; \\ \dfrac{(\delta_{i+2} + \bar{\delta}_{i+2}(1+\tau c))T(c)_{i+1} - T(c)_{i+2}}{1+\tau c}, \ i \leq n-2. \end{cases} \tag{3.3}$$

Here

$$\delta_{i+2} = \frac{T(-1/\tau)_{i+2}}{T(-1/\tau)_{i+1}}, \ i = n-2, n-3, \ldots, 0. \tag{3.4}$$

The *normal conditions* required to complete the sequence are

$$T(-1/\tau)_i \neq 0, \ i = 1, 2, \ldots, n-1. \tag{3.5}$$

Remarks.

1. To determine δ-stability of $F(c)$, one may of course first obtain $\breve{F}(z) = F(c)|_{c \to z}$ and then determine its q-stability by applying familiar stability checking algorithms (e.g., BT or MJT). The possible shortcomings of such a scheme are outlined in [1] and [6]. The purpose here is to obtain a direct check for δ-stability.

2. We follow the work in [6] and [19], and hence, for brevity, all details are omitted.

3. The conditions $T(-1/\tau)_i = 0$, for some $i = 1, 2, \ldots n-1$, imply certain singular conditions on the root distribution of $F(c)$ [11], [19]. The equivalent singular conditions for the real-δ-BT is in [6].

4. Using δ-symmetry, it is easy to show that

$$T(-1/\tau)_i = \frac{\bar{t}_i^{(i)}}{\tau^i}, \ i = 0, 1, \ldots, n. \tag{3.6}$$

Therefore

$$\delta_{i+2} = \frac{1}{\tau} \frac{t_{i+2}^{(i+2)}}{t_{i+1}^{(i+1)}}, \; i = n-2, n-3, \ldots, 0. \tag{3.7}$$

The normal conditions in (3.5) may now be expressed as

$$t_{i+1}^{(i+1)} \neq 0, \; i = n-2, n-3, \ldots, 0. \tag{3.8}$$

Analogous to [6], [11], and [19], we then have

THEOREM 3.1. The polynomial $F(c) \in \Im[c]_n$ is stable iff
I. $t_i^{(i)} \neq 0, \; i = n-1, n-2, \ldots, 1$, and
II. $\nu_n = \text{var}\{T(0)_n, T(0)_{n-1}, \ldots, T(0)_0\} = 0$.

One of the main advantages of the complex-q-BT is that all computations may be carried out through real arithmetic only [19]. The same holds true for the the complex-δ-BT introduced above as well. To see this, let

$$T(c)_i = S(c)_i + jA(c)_i \quad \text{with} \quad \delta_i = \text{Re}[\delta_i] + j\text{Im}[\delta_i], \tag{3.9}$$

for $i = 2, 3, \ldots, n$. It is easy to show that $S(c)_i$'s and $A(c)_i$'s form sequences of δ-symmetric and δ-antisymmetric polynomials, respectively. Now, (3.3) may be expressed as

$$\begin{aligned} S(c)_{i-2} &= \frac{1}{1+\tau c} \left[\text{Re}[\delta_i](2+\tau c) \cdot S(c)_{i-1} + \text{Im}[\delta_i]\tau c \cdot A(c)_{i-1} - S(c)_i\right]; \\ A(c)_{i-2} &= \frac{1}{1+\tau c} \left[-\text{Im}[\delta_i]\tau c \cdot S(c)_{i-1} + (2+\tau c)\text{Re}[\delta_i] \cdot A(c)_{i-1} - A(c)_i\right], \end{aligned} \tag{3.10}$$

for $i = 2, 3, \ldots, n$.

Remark. Note that, $T(0)_i = S(0)_i + jA(0)_i = S(0)_i$.

In the real-δ-BT construction, a certain 'scaling' of $\{T(c)_i\}_{i=0}^n$ was useful [6]. We use the same technique in the complex-δ-BT case as well, thus providing the following advantages: (a) Terms containing τ are avoided during construction, (b) δ_i and ν_i may be deduced by simple inspection, and thus (c) computational effort is reduced.

The sequence of polynomials that incorporates 'scaling' is $\{U(\zeta)_i\}_{i=0}^n$ where

$$U(\zeta)_i = \sum_{k=0}^{i} u_k^{(i)} \zeta^k = T(c)_i \Big|_{c=-\zeta/\tau} \Longleftrightarrow u_k^{(i)} = \left(-\frac{1}{\tau}\right)^k t_k^{(i)},\ k = 0, 1, \ldots, i, \tag{3.11}$$

for $i = 0, 1, \ldots, n$. Thus, from (3.3), we get, for $i = n-2, n-3, \ldots 0$,

$$\begin{aligned} u_0^{(i)} &= (\delta_{i+2} + \bar{\delta}_{i+2}) u_0^{(i+1)} - u_0^{(i+2)}; \\ u_k^{(i)} &= (\delta_{i+2} + \bar{\delta}_{i+2}) u_k^{(i+1)} - \bar{\delta}_{i+2} u_{k-1}^{(i+1)} - u_k^{(i+2)} + u_{k-1}^{(i)},\ k = 1, 2, \ldots i. \end{aligned} \tag{3.12}$$

Note that

$$\delta_{i+2} = \frac{1}{\tau} \frac{\bar{t}_{i+2}^{(i+2)}}{\bar{t}_{i+1}^{(i+1)}} = -\frac{\bar{u}_{i+2}^{(i+2)}}{\bar{u}_{i+1}^{(i+1)}},\ i = n-2, n-3, \ldots, 0, \tag{3.13}$$

and

$$\nu_n = \mathrm{var}\{T(0)_i\}_{i=0}^n = \mathrm{var}\{u_0^{(i)}\}_{i=0}^n. \tag{3.14}$$

Therefore, condition II of Theorem 3.1 may be checked by inspecting the constant coefficients of $\{U(\zeta)_i\}_{i=0}^n$.

Remark. One may use the same 'scaling' strategy in an implementation that uses only real arithmetic.

Relationship between complex-q-BT and complex-δ-BT

As was agreed upon previously, given $F(c)_n \in \Im[z]$, let us use the notation $\breve{F}(z)_n$ to indicate

$$\breve{F}(z)_n = \lambda F(c)_n \Big|_{c \to z} \tag{3.15}$$

where $\lambda \in \Re$ is a possible scaling constant. The establishment of the relationship between the rows of complex-q-BT of $\breve{F}(z)$, i.e., $\{\breve{T}(z)_i\}_{i=0}^n$, and complex-$\delta$-BT of $F(c)$, i.e., $\{T(c)_i\}_{i=0}^n$, which is the subject of this section, is useful later in obtaining the Schur-Cohn minors from the latter.

CLAIM 3.2.

$$\breve{F}^\sharp(z)_n = \lambda F^\sharp(c)_n \Big|_{c \to z}$$

Proof. Note that

$$\breve{F}^{\sharp}(z)_n = z^n \bar{\breve{F}}\left(\frac{1}{z}\right)_n = \lambda z^n \bar{F}(c)_n \Big|_{c\to z} = \lambda z^n \bar{F}\left(\frac{1-z}{\tau z}\right)_n;$$
$$F^{\sharp}(c)_n \Big|_{c\to z} = (1+\tau c)^n \bar{F}\left(-\frac{c}{1+\tau c}\right)_n \Big|_{c\to z} = z^n \bar{F}\left(\frac{1-z}{\tau z}\right)_n.$$

The claim is thus proven. ∎

THEOREM 3.3. The rows of the complex-q-BT of $\breve{F}(z)$ and the complex-δ-BT of $F(c)$ are related by

$$\breve{T}(z)_i = \begin{cases} \lambda T(c)_i \Big|_{c\to z}, & i = n, n-2, \ldots; \\ \frac{\lambda}{\tau} T(c)_i \Big|_{c\to z}, & i = n-1, n-3, \ldots. \end{cases}$$

Proof. First, using Claim 3.2, note that

$$\breve{T}(z)_n = \lambda T(c)_n \Big|_{c\to z}.$$

Thus, Theorem 3.3 is established for $i = n$. $i = n-1$ may also be established directly. For $i = n-2, n-3, \ldots, 0$, use (2.3) and (3.3). ∎

COROLLARY 3.4.

$$\breve{t}_0^{(i)} = \begin{cases} \frac{\lambda}{\tau^i} t_i^{(i)}, & \text{for } i = n, n-2, \ldots; \\ \frac{\lambda}{\tau^{i+1}} t_i^{(i)}, & \text{for } i = n-1, n-3, \ldots, \end{cases}$$
$$\breve{t}_i^{(i)} = \begin{cases} \frac{\lambda}{\tau^i} \bar{t}_i^{(i)}, & \text{for } i = n, n-2, \ldots; \\ \frac{\lambda}{\tau^{i+1}} \bar{t}_i^{(i)}, & \text{for } i = n-1, n-3, \ldots. \end{cases}$$

Proof. This follows directly from Theorem 3.3. ∎

4. Schur-Cohn Minors for δ-Systems

We now develop quantities that may be considered the analogs of Schur-Cohn minors for δ-system polynomials.

LEMMA 4.1. The relationship between the complex-δ-BT of $F(c)_n \in \Im[c]_n$ and the Schur-Cohn minors $\breve{\Delta}_i$, $i = 1, 2, \ldots, n$, of $\breve{F}(z)_n \in \Im[z]_n$ is

$$\breve{\Delta}_i = \frac{\lambda^2}{2\tau^{2(n-i+1)}} \left[(t_{n-i+1}^{(n-i+1)} \bar{t}_{n-i}^{(n-i)} + \bar{t}_{n-i+1}^{(n-i+1)} t_{n-i}^{(n-i)}) \breve{\Delta}_{i-1} \right.$$
$$\left. - \frac{\lambda^2}{2\tau^{2(n-i+1)}} |t_{n-i+1}^{(n-i+1)} t_{n-i}^{(n-i)}|^2 \breve{\Delta}_{i-2} \right], \ \breve{\Delta}_0 = 1, \ \breve{\Delta}_i = 0, \ i < 0.$$

Proof. Note that, the relationship between the complex-q-BT of $\breve{F}(z)_n$ and its Schur-Cohn minors are given by [25]

$$\breve{\Delta}_i = \frac{1}{2} \left[(\breve{t}_0^{(n-i+1)} \breve{t}_{n-i}^{(n-i)} + \breve{t}_{n-i+1}^{(n-i+1)} \breve{t}_0^{(n-i)}) \breve{\Delta}_{i-1} - \frac{1}{2} |\breve{t}_{n-i+1}^{(n-i+1)} \breve{t}_{n-i}^{(n-i)}|^2 \breve{\Delta}_{i-2} \right],$$

with $\breve{\Delta}_0 = 1$ and $\breve{\Delta}_i = 0$, $i < 0$. Now, the claim follows from Corollary 3.4. ∎

Let

$$D = \operatorname{diag} \left\{ \frac{1}{\tau^n}, \frac{1}{\tau^{n-1}}, \ldots, \frac{1}{\tau} \right\} \in \Re^{n \times n}. \tag{4.1}$$

Then, from Lemma 4.1, $\breve{\Delta}$ in Theorem 2.5 is given by

$$\breve{\Delta} = \lambda^2 \cdot D \cdot \Delta \cdot D \tag{4.2}$$

where

$$\Delta = \begin{bmatrix} \operatorname{Re}[t_n^{(n)} \bar{t}_{n-1}^{(n-1)}] & \frac{\tau}{2}[\bar{t}_{n-1}^{(n-1)} t_{n-2}^{(n-2)}] & 0 & \cdots & 0 \\ \frac{\tau}{2}[t_{n-1}^{(n-1)} \bar{t}_{n-2}^{(n-2)}] & \operatorname{Re}[t_{n-1}^{(n-1)} \bar{t}_{n-2}^{(n-2)}] & \frac{\tau}{2}[\bar{t}_{n-2}^{(n-2)} t_{n-3}^{(n-3)}] & \ddots & 0 \\ 0 & \frac{\tau}{2}[t_{n-2}^{(n-2)} \bar{t}_{n-3}^{(n-3)}] & \operatorname{Re}[t_{n-2}^{(n-2)} \bar{t}_{n-3}^{(n-3)}] & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \operatorname{Re}[t_1^{(1)} \bar{t}_0^{(0)}] \end{bmatrix}. \tag{4.3}$$

Clearly, positive definiteness of $\breve{\Delta}$ and Δ are equivalent statements. Hence, we may consider the principal minors of Δ to be the Schur-Cohn minors of $F(c)$.

DEFINITION 4.1. The Schur-Cohn minors of $F(c) \in \Im[c]_n$ are the principal minors of the tridiagonal Hermitian matrix Δ in (4.3).

Therefore, from Theorem 2.4, we have

THEOREM 4.2. The polynomial $F(c) \in \Im[c]_n$ is stable iff $\Delta_i > 0,\ i = 1, 2, \ldots, n$, where Δ_i is the $(i \times i)$-principal minor of Δ in (4.3).

Remarks.

1. Tridiagonal Hermitian matrices constitute an important class of matrices that have been extensively investigated in matrix theory literature [26]. See also [10].
2. Since the Schur-Cohn minor $\breve{\Delta}_i$ obtained from the complex-q-BT are necessarily proper [10], [25], the Schur-Cohn minors defined above for δ-systems are proper as well.

In terms of the 'scaled' sequence of polynomials $\{U(\zeta)_i\}_{i=0}^n$, Theorem 4.2 may be stated as

COROLLARY 4.3. The polynomial $F(c) \in \Im[c]_n$ is stable iff $\tilde{\Delta}_i > 0,\ i = 1, 2, \ldots, n$, where $\tilde{\Delta}_i$ is the $(i \times i)$-principal minor of

$$\tilde{\Delta} = \begin{bmatrix} -\mathrm{Re}[u_n^{(n)}\bar{u}_{n-1}^{(n-1)}] & -\frac{1}{2}[\bar{u}_{n-1}^{(n-1)}u_{n-2}^{(n-2)}] & 0 & \cdots & 0 \\ -\frac{1}{2}[u_{n-1}^{(n-1)}\bar{u}_{n-2}^{(n-2)}] & -\mathrm{Re}[u_{n-1}^{(n-1)}\bar{u}_{n-2}^{(n-2)}] & -\frac{1}{2}[\bar{u}_{n-2}^{(n-2)}u_{n-3}^{(n-3)}] & \ddots & 0 \\ 0 & -\frac{1}{2}[u_{n-2}^{(n-2)}\bar{u}_{n-3}^{(n-3)}] & -\mathrm{Re}[u_{n-2}^{(n-2)}\bar{u}_{n-3}^{(n-3)}] & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & -\mathrm{Re}[u_1^{(1)}\bar{u}_0^{(0)}] \end{bmatrix}. \tag{4.3}$$

Proof. Using (3.11), and factoring out the appropriate diagonal matrices, the result immediately follows. ∎

Remark. Again, notice how the use of the 'scaled' sequence simplifies the entries.

5. Algorithm for Checking Stability of 2-D δ-Systems

To check condition II of Theorem 2.2, we may adopt the following approach:
(a) Express $F(c_1, c_2) \in \Re[c_1]_{n_1}[c_2]_{n_2}$ as a polynomial in $\Im[c_2]_{n_2}$ so that its coefficients, as well as the corresponding Schur-Cohn minors, are parameterized by $c_1 \in \mathcal{T}_\delta$. Here, we have assumed that $n_1 \geq n_2$; otherwise, the roles of n_1 and n_2 may be interchanged.
(b) Check positivity of each of the Schur-Cohn minors, or positive definiteness of the tridiagonal Hermitian matrix $\Delta \in \Im^{n_2 \times n_2}$, for all $c_1 \in \mathcal{T}_\delta$ (see condition II of Theorem 2.2 and Theorem 4.2). These checks may be simplified by applying a direct extension of Siljak's result [16].

However, construction of the complex-δ-BT and the entries of Δ require complex conjugation of certain entries that are functions of $c_1 \in \mathcal{T}_\delta$. This of course complicates the scheme since $\bar{c}_1 = -c_1/(1+\tau c_1)$, $\forall c_1 \in \mathcal{T}_\delta$. On the other hand, in dealing with 2-D q-system stability, we have $\bar{z}_1 = 1/z_1$, $\forall z_1 \in \mathcal{T}_q$. This simple relationship has led to stability checking schemes that use the complex forms of tabular forms [10] that incorporate the *polynomial array* method [27]. To circumvent the above difficulty, the algorithm given below uses the real-δ-BT in order to check Theorem 2.3. In the appendix, an easily implementable algorithm that yields

$$G(x, c_2) = G(x)_{n_1}(c_2)_{2n_2} = F(c_1, c_2)F(\bar{c}_1, c_2)\big|_{\substack{c_1 \in \mathcal{T}_\delta \\ x=(c_1+\bar{c}_1)/2}} \in \Re[x]_{n_1}[c_2]_{2n_2} \tag{5.1}$$

is provided. Note that

$$c_1 \in \mathcal{T}_\delta \iff x \in [-2/\tau, 0]. \tag{5.2}$$

Before proceeding, however, it is important to note that tabular methods are useful in checking for no roots to be *outside* the stability region. However, since in typical 2-D stability studies the 2-D transforms are taken with positive powers [9-10], prior to applying the stability check, the following 'preparation' must be done:
(a) Condition I in Theorem 2.3 may be checked by explicitly finding the roots or applying the real-δ-BT to ensure

$$F^\sharp(c_1)(-1/\tau) \doteq (1+\tau c_1)^{n_1} \bar{F}\left(\frac{-c_1}{1+\tau c_1}\right)\left(-\frac{1}{\tau}\right) \neq 0, \ \forall c_1 \in \Im \setminus \mathcal{U}_\delta \tag{5.3}$$

(that is, polynomial is reciprocated with respect to c_1).

(b) First form

$$G(x)_{n_1}(c_2)_{2n_2} = \sum_{\ell=0}^{2n_2} g_\ell^{(2n_2)}(x) c_2^\ell \in \Re[x]_{n_1}[c_2]_{2n_2} \quad \text{where} \quad g_\ell^{(2n_2)}(x) = \sum_{k=0}^{n_1} g_{k\ell}^{(2n_2)} x^k \in \Re[x]_{n_1}. \tag{5.4}$$

Here $x \in [-2/\tau, 0]$. Now, condition II in Theorem 2.3 may be checked by applying the real-δ-BT to ensure

$$\begin{aligned} \tilde{G}(x)_{n_1}(c_2)_{2n_2} &\doteq \sum_{\ell=0}^{2n_2} \tilde{g}_\ell^{(2n_2)}(x) c_2^\ell \quad \text{where} \quad \tilde{g}_\ell^{(2n_2)}(x) = \sum_{k=0}^{n_1} \tilde{g}_{k\ell}^{(2n_2)} x^k \in \Re[x]_{n_1} \\ &\doteq G(x)^\sharp(c_2) \\ &= (1+\tau c_2)^{2n_2} G(x)\left(\frac{-c_2}{1+\tau c_2}\right) \neq 0, \ \forall x \in [-2/\tau, 0], \ \forall c_2 \in \Im \setminus \mathcal{U}_\delta \end{aligned} \tag{5.5}$$

(that is, polynomial is reciprocated with respect to c_2). Again, $x \in [-2/\tau, 0]$.

We will hence implicitly assume that the given 2-D δ-polynomial has already been appropriately 'prepared' as above. In addition, the construction of the real-δ-BT for $\tilde{G}(x)(c_2)$ requires ensuring [11]

$$\tilde{g}_0^{(2n_2)}(x) \neq 0 \quad \text{and} \quad \tilde{g}_{2n_2}^{(2n_2)} > 0, \quad \forall x \in [-2/\tau, 0]. \tag{5.6}$$

Violation of the first condition in (5.6) is equivalent to

$$F(c_1)(0) = 0 \quad \text{for some} \quad c_1 \in \mathcal{T}_\delta. \tag{5.7}$$

Assuming, with no loss of generality, $\tilde{g}_{2n_2}^{(2n_2)} > 0$ for some $x \in [-2/\tau, 0]$, violation of the second condition in (5.6) is equivalent to

$$F(c_1)(-1/\tau) = 0 \quad \text{for some} \quad c_1 \in \mathcal{T}_\delta. \tag{5.8}$$

Therefore, each of these violations imply instability. Verifying condition (5.7) must be included in the algorithm. Condition (5.8) is automatically verified when condition I in Theorem 2.3 is checked (see (5.3)).

Then, we have the following

THEOREM 5.1. The 2-D δ-system in (2.7) is stable iff

I. $F(c_1)(-1/\tau) \neq 0$, $\forall c_1 \in \overline{\mathcal{U}}_\delta$, and

II. $F(c_1)(0) \neq 0$, $\forall c_1 \in \mathcal{T}_\delta$, and

III. $\Delta_i(0) > 0$, $\forall i = 1, 2, \ldots, 2n_2$, and

IV. $\Delta_{2n_2}(x) > 0$, $\forall x \in [-2/\tau, 0]$, which is satisfied whenever $\Delta_{2n_2}(x) \neq 0$, $\forall x \in [-2/\tau, 0]$, together with condition III.

Here, Δ is the Hermitian matrix mentioned in Theorem 4.2 corresponding to $\tilde{G}(x)(c_2)$ where $x \in [-2/\tau, 0]$.

Conditions I and II in Theorem 5.1 are easy to carry out (they may in fact be verified by explicitly finding the roots). Condition III and IV require construction of the real-δ-BT and the Schur-Cohn minors for which we now develop polynomial arrays [27]. We also provide a scaling scheme so that the numerical reliability of the resulting algorithm is enhanced.

5.1. Polynomial array for entries of real-δ-BT

Express $G(x)(c_2)$ as

$$G(x)(c_2) = \mathbf{x}^{(n_1)^T} \cdot \mathbf{G} \cdot \mathbf{c}_2^{(2n_2)} \tag{5.9}$$

where $\mathbf{x}^{(n_1)} = [x^{n_1}, x^{n_1-1}, \ldots, 1]^T$, $\mathbf{c}_2^{(2n_2)} = [c_2^{2n_2}, c_2^{2n_2-1}, \ldots, 1]^T$, and $\mathbf{G} = \{g_{i,j}\} \in \Re^{(n_1+1)\times(2n_2+1)}$ is the coefficient matrix. Then, it is easy to show that [6]

$$\tilde{G}(x)(c_2) = \mathbf{x}^{(n_1)^T} \cdot \tilde{\mathbf{G}} \cdot \mathbf{c}_2^{(2n_2)} \quad \text{where} \quad \tilde{\mathbf{G}} = \mathbf{G}\boldsymbol{\tau}^{(2n_2)^{-1}}\mathbf{P}^{(2n_2)}\boldsymbol{\tau}^{(2n_2)}. \tag{5.10}$$

Here

$$\begin{aligned} \boldsymbol{\tau}^{(2n_2)} &= \operatorname{diag}\{\tau^{2n_2}, \tau^{2n_2-1}, \ldots, 1\} \in \Re^{(2n_2+1)\times(2n_2+1)}; \\ \mathbf{P}^{(2n_2)} &= \{p_{ij}\} \in \Re^{(2n_2+1)\times(2n_2+1)} \quad \text{where} \quad p_{ij} = (-1)^{2n_2+1-i}\rho_{ij}. \end{aligned} \tag{5.11}$$

The elements ρ_{ij}, which in fact are those of the Pascal's triangle, are given by

$$\rho_{ij} = \begin{cases} 0, & \text{for } i < j; \\ 1, & \text{for } i = j; \\ \rho_{i-1,j-1} + \rho_{i-1,j}, & \text{elsewhere.} \end{cases} \tag{5.12}$$

The real-δ-BT is constructed using the 'scaled' polynomial sequence in (3.11-14). Let

$$\begin{aligned}\tilde{H}(y)(\zeta) &= \tilde{G}(x)(c_2)\,\Big|_{\substack{c_2=-\zeta/\tau\\ x=-y/\tau}};\\ H(y)(\zeta) &= \tilde{G}(x)^{\sharp}(c_2)\,\Big|_{\substack{c_2=-\zeta/\tau\\ x=-y/\tau}} = G(x)(c_2)\,\Big|_{\substack{c_2=-\zeta/\tau\\ x=-y/\tau}}.\end{aligned} \tag{5.13}$$

Note that, $x \in [-2/\tau, 0]$ iff $y \in [0,2]$. Now, using (5.9-12), row $\#2n_2$ and $2n_2-1$ of the corresponding 'scaled' real-δ-BT are given by

$$\begin{aligned}U(y)(\zeta)_{2n_2} &= \sum_{\ell=0}^{2n_2} u_\ell^{(2n_2)}\zeta^\ell = \tilde{H}(y)(\zeta) + H(y)(\zeta)\\ &= \mathbf{y}^{(n_1)^T}\cdot\hat{\boldsymbol{\tau}}^{(n_1)^{-1}}\mathbf{G}\boldsymbol{\tau}^{(2n_2)^{-1}}(\hat{I}^{(2n_2)} + \hat{\mathbf{P}}^{(2n_2)})\cdot\boldsymbol{\zeta}^{(2n_2)};\\ U(y)(\zeta)_{2n_2-1} &= \sum_{\ell=0}^{2n_2-1} u_\ell^{(2n_2-1)}\zeta^\ell = \frac{\tilde{H}(y)(\zeta) - H(y)(\zeta)}{-\zeta/\tau}\\ &= \mathbf{y}^{(n_1)^T}\cdot\hat{\boldsymbol{\tau}}^{(n_1)^{-1}}\tau\mathbf{G}\boldsymbol{\tau}^{(2n_2)^{-1}}(\hat{I}^{(2n_2)} - \hat{\mathbf{P}}^{(2n_2)})\cdot\begin{bmatrix}\boldsymbol{\zeta}^{(2n_2-1)}\\ 0\end{bmatrix},\end{aligned} \tag{5.14}$$

where $\boldsymbol{\zeta}^{(2n_2)} = [\zeta^{2n_2}, \zeta^{2n_2-1}, \ldots, 1]^T$, and

$$\begin{aligned}\hat{\boldsymbol{\tau}}^{(n_1)} &= \mathrm{diag}\{(-\tau)^{n_1}, (-\tau)^{n_1-1}, \ldots, 1\} \in \Re^{(n_1+1)\times(n_1+1)};\\ \hat{I}^{(2n_2)} &= \mathrm{diag}\{(-1)^{2n_2}, (-1)^{2n_2-1}, \ldots, 1\} \in \Re^{(2n_2+1)\times(2n_2+1)};\\ \hat{\mathbf{P}}^{(2n_2)} &= \{\hat{p}_{ij}\} \in \Re^{(2n_2+1)\times(2n_2+1)} \quad \text{where} \quad \hat{p}_{ij} = (-1)^{i+j}\rho_{ij}.\end{aligned} \tag{5.15}$$

Each element of the remaining rows is of the form

$$u_\ell^{(i)}(y) = \frac{n_\ell^{(i)}(y)}{d^{(i)}(y)},\ \ell = 0,1,\ldots,i,\ i = 2n_2, 2n_2-1, \ldots, 0, \tag{5.16}$$

where $n_\ell^{(i)}(y) \in \Re[y]_{\sigma^{(i)}}$ and $d^{(i)}(y) \in \Re[y]_{\varsigma^{(i)}}$. Substituting in (3.12), it is easy to show that, for $\ell = 0,1,\ldots,i$,

$$\begin{aligned}n_\ell^{(i)} &= n_{i+2}^{(i+2)}(n_{\ell-1}^{(i+1)} - 2n_\ell^{(i+1)}) - n_{i+1}^{(i+1)}n_\ell^{(i+2)} + n_{\ell-1}^{(i)}, \quad \text{for} \quad i = 2n_2-2, \ldots, 0;\\ d^{(i)} &= \begin{cases}1, & \text{for } i = 2n_2, 2n_2-1,\\ d^{(i+2)}n_{i+1}^{(i+1)}, & \text{for } i = 2n_2-2, \ldots, 0.\end{cases}\end{aligned} \tag{5.17}$$

Note that $u_\ell^{(2n_2)} = n_\ell^{(2n_2)}$ and $u_\ell^{(2n_2-1)} = n_\ell^{(2n_2-1)}$. Moreover

$$\begin{aligned}\sigma^{(i)} &= \begin{cases}n_1, & \text{for } i = 2n_2, 2n_2-1,\\ \sigma^{(i+2)} + \sigma^{(i+1)}, & \text{for } i = 2n_2-2, \ldots, 0;\end{cases}\\ \varsigma^{(i)} &= \begin{cases}0, & \text{for } i = 2n_2, 2n_2-1,\\ \sigma^{(i)} - n_1, & \text{for } i = 2n_2-2, \ldots, 0.\end{cases}\end{aligned} \tag{5.18}$$

Scaling scheme. Let us scale rows $\#2n_2$ and $\#(2n_2-1)$ so that each coefficient takes values in $[-1,1]$. Correspondingly, for $\ell = 0,1,\ldots,i;\ i = 2n_2, 2n_2-1$, let

$$\begin{aligned} n_\ell^{(i)} &= \lambda^{(i)} \check{n}_\ell^{(i)}; \\ d^{(i)} &= \gamma^{(i)} \check{d}^{(i)}, \end{aligned} \tag{5.19}$$

where $\lambda^{(i)}, \gamma^{(i)} > 0,\ i = 2n_2, 2n_2-1$, are the scaling constants and $[\check{\cdot}]$ denote scaled quantities. Note that

$$\begin{aligned} \frac{n_\ell^{(2n_2)}}{d^{(2n_2)}} &= \frac{\lambda^{(2n_2)}}{\gamma^{(2n_2)}} \frac{\check{n}_\ell^{(2n_2)}}{\check{d}^{(2n_2)}}; \\ \frac{n_\ell^{(2n_2-1)}}{d^{(2n_2-1)}} &= \frac{\lambda^{(2n_2-1)}}{\gamma^{(2n_2-1)}} \frac{\check{n}_\ell^{(2n_2-1)}}{\check{d}^{(2n_2-1)}}. \end{aligned} \tag{5.20}$$

Now, substituting in (5.17-18), we get

$$\begin{aligned} \frac{n_\ell^{(2n_2-2)}}{\lambda^{(2n_2)}\lambda^{(2n_2-1)}} &= \check{n}_{2n_2}^{(2n_2)}(\check{n}_{\ell-1}^{(2n_2-1)} - 2\check{n}_\ell^{(2n_2-1)}) - \check{n}_{2n_2-1}^{(2n_2-1)}\check{n}_\ell^{(2n_2)} + \frac{n_{\ell-1}^{(2n_2-2)}}{\lambda^{(2n_2)}\lambda^{(2n_2-1)}}; \\ \frac{d^{(2n_2-2)}}{\gamma^{(2n_2)}\lambda^{(2n_2-1)}} &= \check{d}^{(2n_2)}\check{n}_{2n_2-1}^{(2n_2-1)}. \end{aligned} \tag{5.21}$$

It can now be seen that, it is only necessary to compute the quantities on the left hand side of (5.21). Then, one may scale these to get

$$\begin{aligned} n_\ell^{(2n_2-2)} &= \lambda^{(2n_2)}\lambda^{(2n_2-1)}\lambda^{(2n_2-2)}\check{n}_\ell^{(2n_2-2)}; \\ d^{(2n_2-2)} &= \gamma^{(2n_2)}\gamma^{(2n_2-2)}\lambda^{(2n_2-1)}\check{d}^{(2n_2-2)}. \end{aligned} \tag{5.22}$$

Note that

$$\frac{n_\ell^{(2n_2-2)}}{d^{(2n_2-2)}} = \frac{\lambda^{(2n_2)}\lambda^{(2n_2-2)}}{\gamma^{(2n_2)}\gamma^{(2n_2-2)}} \frac{\check{n}_\ell^{(2n_2-2)}}{\check{d}^{(2n_2-2)}}. \tag{5.23}$$

Continuing in this manner, the computation of the entries of real-δ-BT may be summarized as follows:

(a) From (5.14), compute $n_\ell^{(i)}, d^{(i)},\ i = 2n_2, 2n_2-1$.

(b) From (5.19), use scaling constants $\lambda^{(i)}, \gamma^{(i)},\ i = 2n_2, 2n_2-1$, to get $\check{n}_\ell^{(i)}, \check{d}^{(i)},\ i = 2n_2, 2n_2-1$.

(c) From (5.21), for $\ell = 0,1,\ldots,i;\ i = 2n_2-2,\ldots,0$, compute

$$\begin{aligned} \frac{n_\ell^{(i)}}{K_n^{(i)}} &= \check{n}_{i+2}^{(i+2)}(\check{n}_{\ell-1}^{(i+1)} - 2\check{n}_\ell^{(i+1)}) - \check{n}_{i+1}^{(i+1)}\check{n}_\ell^{(i+2)} + \frac{n_{\ell-1}^{(i)}}{K_n^{(i)}}; \\ \frac{d^{(i)}}{K_d^{(i)}} &= \check{d}^{(i+2)}\check{n}_{i+1}^{(i+1)}, \end{aligned} \tag{5.24}$$

and use scaling constants $\lambda^{(i)}, \gamma^{(i)}$, $i = 2n_2 - 2, \ldots, 0$, to get $\check{n}_\ell^{(i)}, \check{d}^{(i)}$, $i = 2n_2 - 2, \ldots, 0$. Here, $K_n^{(i)}$ and $K_d^{(i)}$ are constants.

(d) Notice the relationships

$$\frac{n_\ell^{(i)}}{d^{(i)}} = \begin{cases} \frac{\lambda^{(2n_2)}\lambda^{(2n_2-2)}\ldots\lambda^{(i)}}{\gamma^{(2n_2)}\gamma^{(2n_2-2)}\ldots\gamma^{(i)}} \frac{\check{n}_\ell^{(i)}}{\check{d}^{(i)}}, & \text{for } i = 2n_2, 2n_2-2, \ldots, 0; \\ \frac{\lambda^{(2n_2-1)}\lambda^{(2n_2-3)}\ldots\lambda^{(i)}}{\gamma^{(2n_2-1)}\gamma^{(2n_2-3)}\ldots\gamma^{(i)}} \frac{\check{n}_\ell^{(i)}}{\check{d}^{(i)}}, & \text{for } i = 2n_2-1, 2n_2-3, \ldots, 1. \end{cases} \tag{5.25}$$

5.2. Polynomial array for Schur-Cohn minors

Each Schur-Cohn minor obtained from the table, in general, will be of the form

$$\Delta_i(y) = \frac{N^{(i)}(y)}{D^{(i)}(y)} \in \Re(y),\ i = 1, 2, \ldots, 2n_2, \tag{5.26}$$

where $N^{(i)}(y) \in \Re[y]_{\rho^{(i)}}$ and $D^{(i)}(y) \in \Re[y]_{\varrho^{(i)}}$. From Corollary 4.3, we get

$$\Delta_i(y) = -u_{2n_2-i+1}^{(2n_2-i+1)} u_{2n_2-i}^{(2n_2-i)} \Delta_{i-1} - \frac{1}{4} u_{2n_2-i+1}^{(2n_2-i+1)^2} u_{2n_2-i}^{(2n_2-i)^2} \Delta_{i-2},\ i = 1, 2, \ldots, 2n_2, \tag{5.27}$$

where $\Delta_0 \doteq 1$ and $\Delta_i = 0$, $\forall i < 0$.

Remark. Actually, as in [10], one may show that, for stability determination purposes, only the numerator polynomials of Δ_i need be computed. However, to contain the orders of the resulting polynomials, and hence improve numerically conditioning, we do not recommend this scheme.

Scaling scheme. Due to the scaling of entries of the real-δ-BT, computation of Δ_i, $i = 1, 2, \ldots, 2n_2$, may be modified as follows: Let

$$\begin{aligned} \Delta_1 &= -u_{2n_2}^{(2n_2)} u_{2n_2-1}^{(2n_2-1)} \\ &= -\frac{\lambda^{(2n_2)}\lambda^{(2n_2-1)}}{\gamma^{(n2_2)}\gamma^{(2n_2-1)}} \frac{\check{n}_{2n_2}^{(2n_2)} \check{n}_{2n_2-1}^{(2n_2-1)}}{\check{d}^{(2n_2)}\check{d}^{(2n_2-1)}}. \end{aligned} \tag{5.28}$$

Hence, it is only necessary to compute the quantity

$$\check{\Delta}_1 \doteq -\frac{\check{n}_{2n_2}^{(2n_2)} \check{n}_{2n_2-1}^{(2n_2-1)}}{\check{d}^{(2n_2)}\check{d}^{(2n_2-1)}}. \tag{5.29}$$

Continuing in this manner, the computation of the Schur-Cohn minors may be summarized as follows: From (5.27), for $i = 1, 2, \ldots, 2n_2$, compute

$$\check{\Delta}_i = -\frac{\check{n}_{2n_2-i+1}^{(2n_2-i+1)}\check{n}_{2n_2-i}^{(2n_2-i)}}{\check{d}^{(2n_2-i+1)}\check{d}^{(2n_2-i)}}\left[\check{\Delta}_{i-1} + \frac{1}{4}\frac{\lambda^{(2n_2-i)}}{\gamma^{(2n_2-i)}}\frac{\check{n}_{2n_2-i+1}^{(2n_2-i+1)}\check{n}_{2n_2-i}^{(2n_2-i)}}{\check{d}^{(2n_2-i+1)}\check{d}^{(2n_2-i)}}\check{\Delta}_{i-2}\right], \tag{5.30}$$

where $\check{\Delta}_0 \doteq 1$ and $\check{\Delta}_i = 0$, $\forall i < 0$.

Remark. Note that, since $\Delta_i(y)$ is necessarily a proper polynomial (that is, denominator divides numerator properly with no remainder), and not a rational polynomial (see Remark 2 after Theorem 4.2), it is easy to see that $\check{d}^{(2n_2-i+1)}\check{d}^{(2n_2-i)}$ must divide $\check{n}_{2n_2-i+1}^{(2n_2-i+1)}\check{n}_{2n_2-i}^{(2n_2-i)}$ exactly.

5.3. Algorithm

The following result, which is the basis of the stability checking algorithm, is now obvious from [10] and Theorem 5.1:

THEOREM 5.4. The 2-D δ-system in (2.7) is stable iff

I. $F(c_1, -1/\tau) \neq 0$, $\forall c_1 \in \overline{\mathcal{U}}_\delta$, and

II. $F(c_1)(0) \neq 0$, $\forall c_1 \in \mathcal{T}_\delta$, and

III. $\Delta_i(0) > 0$, $\forall i = 1, 2, \ldots, 2n_2$, and

IV. $\Delta_{2n_2}(y) \neq 0$, $\forall y \in [0, 2]$.

The 2-D stability checking algorithm may now be summarized as follows:

GIVEN.

A 2-D δ-polynomial $F(c_1, c_2) \in \Re[c_1]_{n_1}[c_2]_{n_2}$. Without any loss of generality, assume that $n_1 \geq n_2$, and express $F(c_1, c_2)$ as $F(c_1)_{n_1}(c_2)_{n_2}$.

STEP I. Condition I of Theorem 5.4:

Apply an explicit root location procedure. If result is satisfactory, proceed; otherwise, system is unstable.

STEP II. Condition II of Theorem 5.4:

Apply an explicit root location procedure. If result is satisfactory, proceed; otherwise, system is unstable.

STEP III.

Form $G(y)(c_2)$ using the algorithm in the appendix; then form $U(y)(\zeta)_{2n_2}$ and $U(y)(\zeta)_{2n_2-1}$ from (5.14). These yield $n_\ell^{(2n_2)}$ and $n_\ell^{(2n_2-1)}$. Of course, $d^{(2n_2)} = d^{(2n_2-1)} = 1$.

From (5.19), obtain $\check{n}_\ell^{(2n_2)}$, $\check{n}_\ell^{(2n_2-1)}$, and the associated scaling constants $\lambda^{(2n_2)}$ and $\lambda^{(2n_2-1)}$. Of course, $\check{d}^{(2n_2)} = \check{d}^{(2n_2-1)} = 1$ and $\gamma^{(2n_2)} = \gamma^{(2n_2-1)} = 1$.

STEP IV. Condition III of Theorem 5.4:

Form $\check{\Delta}_1(y)$ from (5.30) and check whether $\check{\Delta}_1(0) > 0$.

If result is satisfactory, form $\check{n}_\ell^{(2n_2-2)}$ and $\check{d}^{(2n_2-2)}$ and the associated scaling constants $\lambda^{(2n_2-2)}$ and $\gamma^{(2n_2-2)}$ from (5.24). Form $\check{\Delta}_2(y)$ from (5.30) and check whether $\check{\Delta}_2(0) > 0$. If result is satisfactory, proceed likewise until $\check{\Delta}_{2n_2}(0) > 0$ is checked. Note that, this requires checking of only the constant coefficients. If result is satisfactory, proceed; otherwise, if the check fails at any $i = 1, 2, \ldots, 2n_2$, system is unstable.

STEP V. Condition IV of Theorem 5.4:

Apply an explit root location procedure to check whether $\check{\Delta}_{2n_2}(y) \neq 0$, $\forall y \in [0, 2]$.

Remarks. The possible numerical difficulties that may arise in using explit root location procedures may be avoided as follows: (a) Steps I and II may be verified using the real-δ-BT [6], and (b) step V may be verified by the Sturm sequence method.

6. Example

The stability checking algorithm presented in the previous section is now illustrated through an example. Polynomial entries are denoted using a self-explanatory shorthand notation where the highest degree coefficient is written first. Moreover, only four decimal digital on the mantissa are shown.

Consider the 2-D polynomial

$$F(c_1, c_2) = [\, c_1^2 \quad c_1 \quad 1\,] \begin{bmatrix} 1 & 50 & 740 \\ 52 & 2700 & 38480 \\ 740 & 38480 & 547600 \end{bmatrix} \begin{bmatrix} c_2^2 \\ c_2 \\ 1 \end{bmatrix}$$

with the sampling time $\tau = 0.1$ s.

STEP I. Condition I of Theorem 5.4:

By applying an explicit root location procedure, one can show that

$$F(c_1)(-1/\tau) = 340c_1^2 + 16680c_1 + 236800 \neq 0, \ \forall c_1 \in \overline{\mathcal{U}}_\delta.$$

STEP II. Condition II of Theorem 5.4:

By applying an explicit root location procedure, one can show that

$$F(c_1)(0) = 740c_1^2 + 38480c_1 + 547600 \neq 0, \ \forall c_1 \in \mathcal{T}_\delta.$$

STEP III. Using the algorithm in Appendix, we get

$$G(y)(\zeta) = [\, y^2 \quad y \quad 1\,] \begin{bmatrix} 1.2800e+03 & 1.2992e+05 & 5.1904e+06 & 9.6141e+07 & 7.0093e+08 \\ 5.2480e+04 & 5.4011e+06 & 2.1662e+08 & 3.9968e+09 & 2.8738e+10 \\ 5.4760e+05 & 5.6950e+07 & 2.2912e+09 & 4.2143e+10 & 2.9987e+11 \end{bmatrix} \begin{bmatrix} \zeta^2 \\ \zeta \\ 1 \end{bmatrix}.$$

After scaling, rows #4 and #3 are computed as follows:

$$\check{n}_4^{(4)} = [1.2859e-02, \ -5.0693e-02, \ 5.1315e-02];$$

$$\check{n}_3^{(4)} = [-7.9833e-02, \ 3.1991e-01, \ -3.2798e-01];$$

$$\check{n}_2^{(4)} = [1.9671e-01, \ -7.9909e-01, \ 8.2798e-01];$$

$$\check{n}_1^{(4)} = [-2.3375e-01, \ 9.5836e-01, \ -1.0000e+00];$$

$$\check{n}_0^{(4)} = [1.1687e-01, \ -4.7918e-01, \ 5.0000e-01],$$

with $\lambda^{(4)} = 1.1995e + 12$, and

$$\check{n}_3^{(3)} = [-2.4050e - 02,\ 9.4053e - 02,\ -9.4595e - 02];$$
$$\check{n}_2^{(3)} = [1.3044e - 01,\ -5.1542e - 01,\ 5.2252e - 01];$$
$$\check{n}_1^{(3)} = [-2.4703e - 01,\ 9.8194e - 01,\ -1.0000e + 00];$$
$$\check{n}_0^{(3)} = [1.6469e - 01,\ -6.5463e - 01,\ 6.6667e - 01],$$

with $\lambda^{(3)} = 5.3490e + 10$. Of course, $\check{d}^{(4)} = \check{d}^{(3)} = 1$ with $\gamma^{(4)} = \gamma^{(3)} = 1$.

STEP IV. Condition III of Theorem 5.4:

We get

$$\check{\Delta}_1 = [3.0926e - 04,\ -2.4286e - 03,\ 7.2184e - 03,\ -9.6217e - 03,\ 4.8541e - 03].$$

Clearly, $\check{\Delta}_1(0) = 4.8541e - 03 > 0$.

Now, row #2 is computed as follows:

$$\check{n}_2^{(2)} = [-8.8619e - 03,\ 6.8253e - 02,\ -1.9920e - 01,\ 2.6099e - 01,\ -1.2957e - 01];$$
$$\check{n}_1^{(2)} = [3.3584e - 02,\ -2.5969e - 01,\ 7.6068e - 01,\ -1.0000e + 00,\ 4.9793e - 01];$$
$$\check{n}_0^{(2)} = [-3.3584e - 02,\ 2.5969e - 01,\ -7.6068e - 01,\ 1.0000e + 00,\ -4.9793e - 01],$$

with $\lambda^{(2)} = 4.2420e - 02$. Also,

$$\check{d}^{(2)} = [-2.5424e - 01,\ 9.9428e - 01,\ -1.0000e + 00],$$

with $\gamma^{(2)} = 9.4595e - 02$. We get

$$\check{\Delta}_2 = [1.8046e - 07,\ -2.8190e - 06,\ 1.9343e - 05,\ -7.6148e - 05,\ 1.8810e - 04,$$
$$-2.9857e - 04,\ 2.9737e - 04,\ -1.6992e - 04,\ 4.2654e - 05].$$

Clearly, $\check{\Delta}_2(0) = 4.2654e - 05 > 0$.

Now, row #1 is computed as follows:

$$\check{n}_1^{(1)} = [2.5168e - 03,\ -2.8515e - 02,\ 1.3555e - 01,\ -3.4597e - 01,\ 5.0000e - 01,$$
$$-3.8792e - 01,\ 1.2623e - 01];$$
$$\check{n}_0^{(1)} = [-5.0336e - 03,\ 5.7031e - 02,\ -2.7110e - 01,\ 6.9194e - 01,\ -1.0000e + 00,$$
$$7.7584e - 01,\ -2.5246e - 01],$$

with $\lambda^{(1)} = 3.0980e-02$. Also,

$$\check{d}^{(1)} = [-3.3954e-02,\ 2.6151e-01,\ -7.6322e-01,\ 1.0000e+00,\ -4.9646e-01],$$

with $\gamma^{(1)} = 2.6099e-01$. We get

$$\begin{aligned}\check{\Delta}_3 = [&4.0500e-10,\ -9.3525e-09,\ 9.9260e-08,\ -6.4020e-07,\ 2.7947e-06,\\ &-8.6990e-06,\ 1.9797e-05,\ -3.3188e-05,\ 4.0679e-05,\ -3.5552e-05,\\ &2.1029e-05,\ -7.5594e-06,\ 1.2489e-06].\end{aligned}$$

Clearly, $\check{\Delta}_3(0) = 1.2489e-06 > 0$.

Now, row #0 is computed as follows:

$$\begin{aligned}\check{n}_0^{(0)} = [&-1.0487e-04,\ 1.9379e-03,\ -1.6174e-02,\ 8.0291e-02,\ -2.6251e-01,\\ &5.9070e-01,\ -9.2642e-01,\ 1.0000e+00,\ -7.1104e-01,\ 3.0076e-01,\\ &-5.7473e-02],\end{aligned}$$

with $\lambda^{(0)} = 4.4719e-02$. Also,

$$\begin{aligned}\check{d}^{(0)} = [&-6.7946e-04,\ 1.0355e-02,\ -6.9373e-02,\ 2.6679e-01,\ -6.4420e-01,\\ &1.0000e+00,\ -9.7458e-01,\ 5.4519e-01,\ -1.3404e-01],\end{aligned}$$

with $\gamma^{(0)} = 9.4174e-01$. We get

$$\begin{aligned}\check{\Delta}_4 = [&4.3531e-12,\ -1.3058e-10,\ 1.8400e-09,\ -1.6166e-08,\ 9.9118e-08,\\ &-4.4970e-07,\ 1.5618e-06,\ -4.2352e-06,\ 9.0628e-06,\ -1.5355e-05,\\ &2.0530e-05,\ -2.1433e-05,\ 1.7129e-05,\ -1.0130e-05,\ 4.1814e-06,\\ &-1.0762e-06,\ 1.3014e-07].\end{aligned}$$

Clearly, $\check{\Delta}_4(0) = 1.3014e-07 > 0$.

STEP V. Condition IV of Theorem 5.4:

By applying an explicit root location procedure, one can show that

$$\check{\Delta}_4(y) \neq 0,\ \forall y \in [0, 2].$$

Thus, we conclude that $F(c_1, c_2)$ is stable.

7. Conclusion and Final Remarks

In this paper, we have developed an efficient stability checking algorithm applicable for 2-D δ-system characteristic polynomials. Our purpose here is to obtain a direct algorithm due to the possible numerical disadvantages associated with indirect methods that utilize transformation techniques.

In arriving at the algorithm, the following contributions have been made: (a) Tabular method of stability checking applicable for δ-system polynomials possibly possessing complex-valued coefficients, (b) quantities that may be regarded as the Schur-Cohn minors applicable for such systems, and (c) polynomial arrays for computing both table entries and Schur-Cohn minors.

The proposed Schur-Cohn minors lets one use a Siljak-like simplification [16] in the stability check. Although the algorithm utilizes only the real-δ-BT, results regarding the Schur-Cohn minors are in fact valid for the more general complex-valued coefficient case as well.

As in [10], it is possible to develop the algorithm such that only the numerator polynomials of the entries of the real-δ-BT and the Schur-Cohn minors are computed. Then, we do not require polynomial division operations. However, our experience has been that such a scheme is prone to be numerically unreliable. This is mainly due to the explosion of polynomial degree especially in computing the Schur-Cohn minors. To avoid these difficulties and enhance numerical reliability, we have (a) introduced a scaling scheme, and (b) used polynomial division to contain the polynomial degree. The latter is not new; in fact, MJT also uses this. If the user is interested in implementing the algorithm using PRO-MATLAB [28], these polynomial division operations may be conveniently performed using the routine **deconv**.

We believe that a suitable scaling strategy can improve the numerical reliability of the MJT as well. The authors are currently looking into this.

The algorithm developed is easily implementable on a computer. The authors have

implemented it via a C-language routine that the interested reader may request from the second author.

Acknowledgement

The authors are indebted to Professor Eliahu I. Jury for many helpful discussions and the reviewers for their careful examination of the manuscript and constructive suggestions. The first author's research work was partially supported by the Office of Naval Research (ONR) through the grant N00014-94-1-0454. This support is gratefully acknowledged.

References

1. R.H. Middleton and G.C. Goodwin, *Digital Control and Estimation: A Unified Approach,* Englewood Cliffs, NJ: Prentice-Hall, 1990.
2. R. Vijayan, H.V. Poor, J.B. Moore, and G.C. Goodwin, "A Levinson-type algorithm for modeling fast-sampled data," *IEEE Transactions on Automatic Control,* vol. 36, pp. 314-321, Mar. 1991.
3. C.B. Soh, "Robust stability of discrete-time systems using delta operators," *IEEE Transactions on Automatic Control,* vol. 36, pp. 377-380, 1991.
4. G.C. Goodwin, R.H. Middleton, and H.V. Poor, "High-speed digital signal processing and control," *Proceedings of the IEEE,* vol. 80, pp. 240-259, 1992.
5. G. Li and M. Gevers, "Roundoff noise minimization using delta-operator realizations," *IEEE Transactions on Signal Processing,* vol. 41, pp. 629-637, 1993.
6. K. Premaratne and E.I. Jury, "Tabular method for determining root distribution of delta operator formulated polynomials," *IEEE Transactions on Automatic Control,* vol. 39, pp. 352-355, 1994.
7. K. Premaratne, R. Salvi, N.R. Habib, and J.P. Le Gall, "Delta-operator formulated discrete-time equivalents of continuous-time systems," *IEEE Transactions on Automatic Control,* vol. 39, pp. 581-585, 1994.
8. G. Likourezos, "Prolog to 'High-speed digital signal processing and control'," *Proceedings of the IEEE,* vol. 80, 238-239, 1992.
9. E.I. Jury, "Stability of multidimensional systems and other related problems," Chapter 3 in *Multidimensional Systems, Techniques, and Applications,* New York, NY: Marcel Dekkar, 1986.
10. K. Premaratne, "Stability determination of two-dimensional discrete-time systems," *Multidimensional Systems and Signal Processing,* vol. 4, pp. 331-354, 1993.
11. Y. Bistritz, "Zero location with respect to the unit circle of discrete-time linear system polynomials," *Proceedings of the IEEE,* vol. 72, pp. 1131-1142, 1984.
12. M. Mansour, "Stability and robust stability of discrete-time systems in the δ-transform," *Fundamentals of Discrete-Time Systems: A Tribute to Professor Eliahu I. Jury,* pp. 133-140, Albuquerque, NM: TSI Press, 1993.
13. R. DeCarlo, J. Murray, and R. Saeks, "Multivariable Nyquist theory," *International Journal of Control,* vol. 25, pp. 657-675, 1977.
14. G. Gu and E.B. Lee, "A numerical algorithm for stability testing of 2-D recursive digital filters," *IEEE Transactions on Circuits and Systems,* vol. 37, pp. 135-138, 1990.
15. E.I. Jury, "Modified stability table for 2-D digital filters," *IEEE Transactions on Circuits and Systems,* vol. 35, pp. 116-119, 1988; "Addendum to 'Modified stability table for 2-D digital filters'," Department Electrical and Computer Engineering, University of Miami, Coral Gables, FL, 1987.
16. D.D. Siljak, "Stability criteria for two-dimensional polynomials," *IEEE Transactions on Circuits and Systems,* vol. 22, pp. 185-189, 1975.
17. E.I. Jury, "A note on the modified stability table for linear discrete-time systems," *IEEE Transactions on Circuits and Systems,* vol. 38, pp. 221-223, 1991.
18. E.I. Jury, *Theory and Application of the z-Transform Method,* John Wiley & Sons: New York, NY, 1964.
19. Y. Bistritz, "A circular stability test for general polynomials," *Systems and Control Letters,* vol. 7, pp. 89-97, 1986.
20. T.S. Huang, "Stability of two-dimensional recursive filters," *IEEE Transactions on Automatic Control,* vol. 20, pp. 158-183, 1973.
21. N.K. Bose, "Implementation of a new stability test for two-dimensional filters," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. 25, pp. 117-120, 1977.
22. B.M. Karan and M.C. Srivastava, "A new stability test for 2-D filters," *IEEE Transactions on Circuits and Systems,* vol. 33, pp. 807-809, 1986.
23. A. Cohn, "Über die Anzahl der Wurzeln einer algebraischen Gleichung in einem Kreise," *Math. Z.,* vol. 14-15, pp. 110-148, 1914.
24. I. Schur, "Über Potenzreihen die in Inners des Einheitskreises besehränkt sind," *J. Für Math.,* vol. 147, pp. 205-232, 1917.
25. K. Premaratne and E.I. Jury, "On the Bistritz tabular form and its relationship with the Schur-Cohn minors and inner determinants," *Journal of the Franklin Institute,* vol. 330, pp. 165-182, 1993.
26. G.H. Golub and C.F. Van Loan, *Matrix Computations,* Baltimore, MD: John Hopkins University Press, 1983.
27. X. Hu and E.I. Jury, "On two-dimensional filter stability test," to appear in *IEEE Transactions on Circuits and Systems,* 1993.
28. *PRO-MATLAB User's Guide,* South Natick, MA: The MathWorks Inc., 1991.

Appendix. Algorithm to obtain $G(x)_{n_1}(c_2)_{2n_2}$ from $F(c_1)_{n_1}(c_2)_{n_2}$

Given

$$F(c_1)_{n_1}(c_2)_{n_2} = \sum_{\ell=0}^{n_2} f_\ell(c_1)\cdot c_2^\ell \quad \text{where} \quad f_\ell(c_1) = \sum_{k=0}^{n_1} f_{k,\ell}\cdot c_1^k,\ c_1\in\mathcal{T}_\delta, \tag{a.1}$$

we now develop an algorithm that yields

$$G(x)_{n_1}(c_2)_{2n_2} \doteq \sum_{j=0}^{2n_2} g_j(x)\cdot c_2^j = F(c_1)_{n_1}(c_2)_{n_2}\cdot F(\bar{c}_1)_{n_1}(c_2)_{n_2},\ c_1\in\mathcal{T}_\delta. \tag{a.2}$$

First, we see

$$\begin{aligned} G(x)(c_2) &= \sum_{\ell=0}^{n_2}\sum_{j=0}^{n_2} f_\ell(c_1)f_j(\bar{c}_1)\cdot c_2^{\ell+j} \\ &= \sum_{\ell=0}^{n_2}\sum_{j=\ell}^{n_2+\ell} f_\ell(c_1)f_{j-\ell}(\bar{c}_1)\cdot c_2^j = \sum_{j=0}^{2n_2}\sum_{\ell=0}^{j} f_\ell(c_1)f_{j-\ell}(\bar{c}_1)\cdot c_2^j \end{aligned} \tag{a.3}$$

(quantities with negative subscripts are taken to be zero). Hence, comparing (a.2-3), we get

$$\begin{aligned} g_j(x) &= \sum_{\ell=0}^{j} f_\ell(c_1)f_{j-\ell}(\bar{c}_1) = \sum_{\ell=0}^{j}\left[\sum_{k=0}^{n_1}\sum_{i=0}^{n_1} f_{k,\ell}f_{i,j-\ell}\cdot c_1^k\bar{c}_1^i\right] \\ &= \sum_{\ell=0}^{j}\left[\sum_{k=0}^{n_1} f_{k,\ell}f_{k,j-\ell}\cdot(c_1\bar{c}_1)^k + \sum_{k=0}^{n_1}\sum_{\substack{i=0\\ i\neq k}}^{n_1} f_{k,\ell}f_{i,j-\ell}\cdot c_1^k\bar{c}_1^i\right] \\ &= \sum_{\ell=0}^{j}\sum_{k=0}^{n_1} f_{k,\ell}f_{k,j-\ell}\cdot(c_1\bar{c}_1)^k + X \end{aligned} \tag{a.4}$$

where

$$\begin{aligned} X &= \sum_{\ell=0}^{j}\sum_{k=0}^{n_1}\left[\sum_{\substack{i=0\\ i\neq k}}^{k} f_{k,\ell}f_{i,j-\ell}\cdot c_1^k\bar{c}_1^i + \sum_{\substack{i=k\\ i\neq k}}^{n_1} f_{k,\ell}f_{i,j-\ell}\cdot c_1^k\bar{c}_1^i\right] \\ &= \sum_{\ell=0}^{j}\left[\sum_{k=0}^{n_1}\sum_{\substack{i=0\\ i\neq k}}^{k} f_{k,\ell}f_{i,j-\ell}\cdot c_1^k\bar{c}_1^i + \sum_{k=0}^{n_1}\sum_{\substack{i=0\\ i\neq k}}^{k} f_{i,\ell}f_{k,j-\ell}\cdot c_1^i\bar{c}_1^k\right] \\ &= \sum_{\ell=0}^{j}\left[\sum_{k=0}^{n_1}\sum_{\substack{i=0\\ i\neq k}}^{k}\left(f_{k,\ell}f_{i,j-\ell}\cdot c_1^{k-i} + f_{i,\ell}f_{k,j-\ell}\cdot\bar{c}_1^{k-i}\right)\cdot(c_1\bar{c}_1)^i\right] \\ &= \sum_{k=0}^{n_1}\sum_{i=0}^{k}(c_1\bar{c}_1)^i\sum_{\ell=0}^{j}\left[f_{k,\ell}f_{i,j-\ell}\cdot c_1^{k-i} + f_{i,j-\ell}f_{k,\ell}\cdot\bar{c}_1^{k-i}\right]. \end{aligned} \tag{a.5}$$

Let us use the notation

$$c_1^{(n)} = \frac{c_1^n + \bar{c}_1^n}{2},\ c_1 \in \mathcal{T}_\delta,\ n = 0, 1, \ldots. \tag{a.6}$$

Noting that, for $c_1 \in \mathcal{T}_\delta$,

$$\bar{c}_1 = -\frac{c_1}{1 + \tau c_1}, \tag{a.7}$$

it is easy to show that

$$c_1 \bar{c}_1 = -\frac{2}{\tau} c_1^{(1)}. \tag{a.8}$$

Substituting in (a.5), we get

$$X = \sum_{\ell=0}^{j} \left[\sum_{k=0}^{n_1} \sum_{i=0}^{k-1} 2 f_{k,\ell} f_{i,j-\ell} \left(\frac{-2}{\tau} \right)^i \cdot c_1^{(1)^i} c_1^{(k-i)} \right]. \tag{a.9}$$

Substituting in (a.4), we get

$$g_j(x) = \sum_{\ell=0}^{j} \sum_{k=0}^{n_1} \left[f_{k,\ell} f_{k,j-\ell} \left(\frac{-2}{\tau} \right)^k \cdot c_1^{(1)^k} + \sum_{i=0}^{k-1} 2 f_{k,\ell} f_{i,j-\ell} \left(\frac{-2}{\tau} \right)^i \cdot c_1^{(1)^i} c_1^{(k-i)} \right]. \tag{a.10}$$

Now, in order to develop the algorithm, we need a recursive procedure to compute $c_1^{(n)}$, $n = 0, 1, \ldots$. To proceed, we note that

$$\begin{aligned} c_1^{(n)} &= \frac{(c_1 + \bar{c}_1)(c_1^{n-1} + \bar{c}_1^{n-1}) - c_1 \bar{c}_1 (c_1^{n-2} + \bar{c}_1^{n-2})}{2} \\ &= 2c_1^{(1)} \left(c_1^{(n-1)} + \frac{1}{\tau} c_1^{(n-2)} \right),\ n = 2, 3, \ldots. \end{aligned} \tag{a.11}$$

Let

$$c_1^{(n)} = \sum_{i=0}^{n} c_{1,i}^{(n)} x^i \tag{a.12}$$

where

$$c_1^{(1)} \doteq x. \tag{a.13}$$

Remark. Note that

$$c_1^{(0)} = 1. \tag{a.14}$$

Substituting (a.12) in (a.11), and equating similar coefficients, we get

$$c_{1,i}^{(n)} = 2 \left(c_{1,i-1}^{(n-1)} + \frac{1}{\tau} c_{1,i-1}^{(n-2)} \right),\ i = 0, \ldots, n,\ n = 2, 3, \ldots. \tag{a.15}$$

For instance, $c_1^{(n)}$, $n = 0, 1, \ldots, 5$, may be conveniently obtained from

$$\begin{bmatrix} c_1^{(0)} \\ c_1^{(1)} \\ c_1^{(2)} \\ c_1^{(3)} \\ c_1^{(4)} \\ c_1^{(5)} \end{bmatrix} = \begin{bmatrix} 1 & & & & & \\ 0 & 1 & & & & \\ 0 & 2/\tau & 2 & & & \\ 0 & 0 & 6/\tau & 4 & & \\ 0 & 0 & 4/\tau^2 & 16/\tau & 8 & \\ 0 & 0 & 0 & 20/\tau^2 & 40/\tau & 16 \end{bmatrix} \begin{bmatrix} 1 \\ x \\ x^2 \\ x^3 \\ x^4 \\ x^5 \end{bmatrix}.$$

IEEE Transactions on Signal Processing, in review, 1995.

SP EDICS NUMBER : **SP 2.1.6**

An Exhaustive Search Algorithm For Checking Limit Cycle Behavior Of Digital Filters

K. Premaratne, *Senior Member, IEEE*, E.C. Kulasekere [1]

P.H. Bauer[2], *Member, IEEE*, L.J. Leclerc [3]

Abstract :-*The presence of limit cycles that may arise in fixed-point arithmetic implementation of a digital filter can significantly impair its performance. The work in this paper presents an algorithm that can be utilized to determine the presence or the absence of such limit cycles of a given digital filter. The filter is assumed to be in its state-space formulation and hence, performance of the corresponding direct form representation follows as a special case. Moreover, the algorithm is applicable independent of the filter order, type of quantization nonlinearity, and whether the accumulator is single-length or double-length. In developing the algorithm, bounds on the amplitude and period of limit cycles of a given digital filter are obtained. The robustness of the algorithm in terms of limit cycles performance with respect to filter coefficient perturbations is verified. Hence, it may be utilized to obtain regions in the coefficient space where a digital filter of given order is limit cycle free.*

[1]K. Premaratne and E.C. Kulasekere are with the Department of Electrical and Computer Engineering, P.O. Box 248294, University of Miami, Coral Gables, FL 33124, USA.

[2]P.H. Bauer is with the Department of Electrical Engineering, Laboratory for Image and Signal Analysis (LISA), University of Notre Dame,Notre Dame, IN 46556, USA.

[3]L.J. Leclerc is with the Ericsson Communications, Inc., Ville Mont-Royal, Québec, CANADA.

K.P. and P.H.B. gratefully acknowledge the support received from the Office of Naval Research (ONR) through the grants N00014-94-1-0454 and N00014-94-1-0387, respectively.

Corresponding Author:
E.C. Kulasekere, Phone : (305) - 284 - 4046, FAX : (305) - 284 - 4044
Email : ekulasek@obsidian.eng.miami.edu

I Introduction

A digital filter may be realized using either a general purpose digital computer or special purpose digital hardware. In either case, the coefficients and intermediate results of computations must be stored in binary form in registers of finite wordlength. Limit cycle oscillations are a direct result of this limitation, and care must be taken to suppress them while performing a digital filter design.

For the past several years, this in fact has been a research topic of interest, and a significant amount of insight and research results are now available [1]-[10]. In an implementation of a higher order digital filter, as shown in [11], a cascade or parallel form composed of first-order and second-order subfilters is preferable over any direct form realization. Therefore the results are summarized for the second order realizations. Most existing results focus on the effects of signed magnitude rounding and truncation quantization schemes with regard to the existence of limit cycles. Recently, some work addressing the two's complement truncation scheme has also appeared [12]-[14].

This work proposes an algorithm that may be used to check for limit cycles of a given digital filter implemented using fixed-point arithmetic. It possesses a wide scope of applicability: The digital filter to be tested may be of any order; the quantization scheme may be arbitrary, including truncation and rounding schemes corresponding to signed magnitude and two's complement; and the accumulator may be of single- or double-length.

Given a digital filter, we develop bounds on the amplitude and period of possible limit cycles. The algorithm is based on an exhaustive search procedure over all these possibilities. In addition, extending the same procedure to the entire linear stability region, one may utilize it to obtain regions in the filter coefficient space where the given filter is globally asymptotically stable (g.a.s.). For this purpose, the robustness of the algorithm in terms of presence or absence of limit cycles with respect to filter coefficient perturbations is also verified. A similar concept

has been used before for checking limit cycle behavior of digital filters implemented in direct form [10], [15]-[18]. The major advantage of the proposed method is that it is applicable for the more general state-space implementations. Of course, the direct form implementation then follows as a special case.

The paper is organized as follows. Section II contains the nomenclature used throughout the paper. Section III provides bounds on the amplitude and period of limit cycles of a given general digital filter. Section IV discusses the algorithm and its computational aspects. Section V addresses the robustness of the algorithm with respect to perturbations of filter coefficients. Section VI contains some situations where the algorithm developed has been used effectively. Finally, Section VII contains the concluding remarks.

II Nomenclature

The following notation will be used throughout the paper.

$\Re$, $\mathcal{Z}$	Set of reals, set of integers.
$\mathcal{C}$	Set of complex numbers.
$\mathcal{Z}_+$	Nonnegative integers.
$\Re^{m\times n}$, $\mathcal{Z}^{m\times n}$	Set of matrices of size $m \times n$ over the reals and integers.
$\Re(z)_{m\times n}$	Set of matrices of size $m \times n$ over the rational polynomials in the indeterminate $z \in \mathcal{C}$.
$\mathcal{K}[\cdot]$	Cardinality of set $[\cdot]$.
a_{ij}	(i,j)-th element of the matrix $A = \{a_{ij}\}$.
I, $\mathbf{0}$	Identity matrix and null matrix of appropriate sizes.
$\mathbf{x}(k)$	Filter state vector at instant k.
$x_i(k)$	i-th component of the state vector $\mathbf{x}(k)$.
$\|\cdot\|_\infty$	The infinity norm. For $\mathbf{x} = \{x_i\} \in \Re^m$, $\|\mathbf{x}\|_\infty = \max_i \|x_i\|$; for $A = \{a_{ij}\} \in \Re^{m\times n}$, $\|A\|_\infty = \max_i \sum_{j=1}^n \|a_{ij}\|$.

M_i	Upper bound for absolute value of amplitude of $x_i(k)$, $k \in \mathcal{Z}_+$.
$\hat{M}_i$	Largest integer less than or equal to M_i.
$\delta(k)$	Dirac delta function.
$H_{ij}(z)$	(i,j)-th element, that is, the (i,j)-th transfer function, of the MIMO transfer function $H(z)$.
$h_{ij}(k)$	Impulse response of $H_{ij}(z)$.
P_i	i-th pole (accounting for multiplicity) of $H_{ij}(z)$.
K_{ij}	Constant term in the partial fraction expansion of $H_{ij}(z)$.
r_{ij}^k	k-th residue of $H_{ij}(z)$.
q	Quantization step size.
$\mathcal{Q}[\cdot]$	Quantization nonlinearity operator.
ϱ	Normalized quantization error. For instance, for roundoff, $\varrho = 0.5$, and for truncation, $\varrho = 1$.
N	Number of nonlinearities in a realization.
$\mathrm{e}(k)$	Quantization error vector.
T	Limit cycle period.
$\mathcal{S}^{(0)}$	Set of state vectors satisfying the upper bound $\hat{M}_i$ such that $\lvert x_i \rvert \leq \hat{M}_i \; \forall i$.
Δa_{ij}	Perturbation of the coefficient a_{ij}.

III Amplitude and Period Bounds on Limit Cycles.

In general, the quantization nonlinearity satisfies

$$|x - \mathcal{Q}[x]| \leq \varrho \cdot q, \quad \forall x \in \Re \tag{1}$$

where ϱ is the normalized quantization error. In particular, for roundoff quantization, $\varrho = 0.5$, and for truncation quantization, $\varrho = 1$. Note that, all the filter parameters may be expressed as integer multiples of the quantization step size q. Hence, for convenience, we normalize q to unity

for all calculations. The quantization nonlinearity thus becomes an integer valued function, viz.,

$$\mathcal{Q} : \Re \rightarrow \mathcal{Z} \tag{2}$$

In general, for all quantization schemes of interest, $\mathcal{Q}[0] = 0$.

We consider a digital filter of order m in its minimal state-space representation $\{A, B, C, D\}$, that is,

$$\mathbf{x}(k+1) = A \cdot \mathbf{x}(k) + B \cdot \mathbf{u}(k); \tag{3}$$

$$\mathbf{y}(k) = C \cdot \mathbf{x}(k) + D \cdot \mathbf{u}(k), \tag{4}$$

where $\mathbf{x} \in \Re^m$ is the state, $\mathbf{u}$ is the input, and $\mathbf{y}$ is the output. Also, $A \in \Re^{m \times m}$. For addressing limit cycle performance, we consider the zero input recursive state equation

$$\mathbf{x}(k+1) = A \cdot \mathbf{x}(k). \tag{5}$$

Unless otherwise stated, we only consider linearly stable filters. Hence, all eigenvalues of A are inside the unit circle in $\mathcal{C}$.

Now, under finite wordlength conditions, the appearance of the pertinent quantization nonlinearity in (5) may be modeled as

$$\mathbf{x}(k+1) = \mathcal{Q}[A \cdot \mathbf{x}(k)]. \tag{6}$$

Depending on whether the result of a product can be stored with full precision or whether quantization is performed immediately after each product is computed determines the effect of this nonlinearity. Considering (5) and noting that $\mathbf{x}(k) = \{x_i\} \in \Re^m$ and $A = \{a_{ij}\} \in \Re^{m \times m}$, we get the following:

If the products can be stored with full precision, that is, if a double-length accumulator is available,

$$\mathbf{x}(k+1) = \begin{pmatrix} \mathcal{Q}[\sum_{j=1}^{m} a_{1j} \cdot x_j(k)] \\ \vdots \\ \mathcal{Q}[\sum_{j=1}^{m} a_{mj} \cdot x_j(k)] \end{pmatrix} \tag{7}$$

and, on the other hand, if the product is quantized immediately after each product is performed, that is, if only a single-length accumulator is available,

$$\mathbf{x}(k+1) = \begin{pmatrix} \mathcal{Q}[a_{11} \cdot x_1(k)] + \mathcal{Q}[a_{12} \cdot x_2(k)] + \ldots + \mathcal{Q}[a_{1m} \cdot x_m(k)] \\ \vdots \\ \mathcal{Q}[a_{m1} \cdot x_1(k)] + \mathcal{Q}[a_{m2} \cdot x_2(k)] + \ldots + \mathcal{Q}[a_{mm} \cdot x_m(k)] \end{pmatrix} \tag{8}$$

Since q has been normalized to unity, noting (1), (7) and (8) may be expressed in a unified manner as

$$\mathbf{x}(k+1) = A \cdot \mathbf{x}(k) + \mathbf{e}(k), \quad \text{with} \quad |e_i(k)| \leq N \cdot \varrho, \tag{9}$$

where $\mathbf{e}(k) = \{e_i(k)\} \in \Re^m$ and $e_i(k) \in \Re$. Note that, if (7) is applicable, $N = 1$; if (8) is applicable, $N = m$.

We note that, (9) is a description of a *linear* system driven by the bounded quantization error input $\mathbf{e}(k)$. Hence, we have in fact converted the nonlinear systems in (7) and (8) into the linear system in (9). Now, the transfer function between $\mathbf{e}(k)$ and $\mathbf{x}(k)$ is

$$\frac{\mathbf{X}(z)}{\mathbf{E}(z)} = (z \cdot I - A)^{-1} \in \Re(z)_{m \times m}, \tag{10}$$

where $\mathbf{X}$ and $\mathbf{E}$ are the z-transforms of $\mathbf{x}$ and $\mathbf{e}$, respectively. This, when expanded, may be expressed as

$$\frac{\mathbf{X}(z)}{\mathbf{E}(z)} = \begin{pmatrix} H_{11}(z) & H_{12}(z) & \ldots & H_{1m}(z) \\ \vdots & \vdots & \ddots & \vdots \\ H_{m1}(z) & H_{m2}(z) & \ldots & H_{mm}(z) \end{pmatrix}$$

where $H_{ij}(z) \in \Re(z)$. Hence,

$$X_i(z) = \sum_{j=1}^{m} H_{ij}(z) \cdot E_j(z), \; i = 1, 2, \ldots, m, \tag{11}$$

where $\mathbf{X}(z) = \{X_i\}$ and $\mathbf{E}(z) = \{E_j\}$. Taking inverse z-transform of the above, we get

$$x_i(k) = \sum_{j=1}^{m} h_{ij}(k) * e_j(k), \; i = 1, 2, \ldots, m,$$

where $h_{ij}(k)$ is the impulse response of $H_{ij}(z)$. Hence

$$x_i(k) = \sum_{j=1}^{m} \sum_{r=0}^{\infty} h_{ij}(r) \cdot e_j(k-r), \; i = 1, 2, \ldots, m. \tag{12}$$

Combining (12) with the fact that $|e_j(k)| \leq N \cdot \varrho$, for $j = 1, 2, \ldots, m$, we obtain

$$|x_i(k)| \leq N \cdot \varrho \cdot \sum_{j=1}^{m} \sum_{k=0}^{\infty} |h_{ij}(k)|. \tag{13}$$

Eqn. (13) may now be used to provide upper bounds for each state vector x_i as follows:

$$M_i = N \cdot \varrho \cdot \sum_{j=1}^{m} \sum_{k=0}^{\infty} |h_{ij}(k)|, \; i = 1, 2, \ldots, m. \tag{14}$$

We realize that, in order to estimate a useful upper bound for each x_i, we need to compute $\sum_{j=1}^{m} \sum_{k=0}^{\infty} |h_{ij}(k)|$ for a given filter. We address this now. Consider the transfer function $H_{ij}(z)$.

All poles of $H_{ij}(z)$ are distinct:

In this case, $H_{ij}(z)$ may be expressed as

$$H_{ij}(z) = K_{ij} + \frac{r_{ij}^{(1)}}{1 - P_1^{(1)} z^{-1}} + \ldots + \frac{r_{ij}^{(m)}}{1 - P_m^{(m)} z^{-1}},$$

where $r_{ij}^{(p)}, P_{\ell}^{(q)} \in \mathcal{C}$ and $K_{ij} \in \Re$, for $i, j, \ell, p, q = 1, 2, \ldots, m$. Taking the inverse z-transform, we have

$$h_{ij}(k) = K_{ij} \cdot \delta(k) + r_{ij}^{(1)} [P_1^{(1)}]^k + \ldots + r_{ij}^{(m)} [P_m^{(m)}]^k,$$

where $\delta(k)$ is the Dirac delta function. Therefore

$$\begin{aligned} \sum_{k=0}^{\infty} |h_{ij}(k)| &\leq \sum_{k=0}^{\infty} \{|K_{ij}||\delta(k)| + |r_{ij}^{(1)}|[|P_1^{(1)}|]^k + \ldots + |r_{ij}^{(m)}|[|P_m^{(m)}|]^k\} \\ &= |K_{ij}| + |r_{ij}^{(1)}|(1 - |P_1^{(1)}|)^{-1} + \ldots + |r_{ij}^{(m)}|(1 - |P_m^{(m)}|)^{-1}. \end{aligned}$$

This, when expanded, gives

$$\begin{aligned} \sum_{j=1}^{m} \sum_{k=0}^{\infty} |h_{ij}(k)| \leq \; & \textstyle\sum_{j=1}^{m} |K_{ij}| \; + (1 - |P_1^{(1)}|)^{-1} \cdot \sum_{j=1}^{m} |r_{ij}^{(1)}| + \ldots \\ & + \ldots \qquad + (1 - |P_m^{(m)}|)^{-1} \cdot \sum_{j=1}^{m} |r_{ij}^{(m)}|, \end{aligned}$$

for $i = 1, 2, \ldots, m$. Hence

$$\begin{aligned} |x_i(k)| \leq \; & N \cdot \varrho \cdot \{\textstyle\sum_{j=1}^{m} |K_{ij}| + (1 - |P_1^{(1)}|)^{-1} \cdot \sum_{j=1}^{m} |r_{ij}^{(1)}| + \ldots \\ & \ldots + (1 - |P_m^{(m)}|)^{-1} \cdot \textstyle\sum_{j=1}^{m} |r_{ij}^{(m)}|\}, \end{aligned} \tag{15}$$

for $i = 1, 2, \ldots, m$. Note that, convergence of the above is guaranteed due to linear stability of the digital filter.

Remark. The method adopted in [10] tends to be easier to implement and more general with regards to its capability of handling the presence of poles of higher multiplicity. However, our experience has been that the technique described above often leads to lower upper bounds. Note that, the technique in [10] utilizes an interpretation that involves a cascade of first-order sections to obtain a bound for $|x_i|$; the technique above utilizes an interpretation that involves a parallel combination. Of course, no *one* technique will provide a lower bound for *all* situations. If computer cost is of concern, one can run both techniques and utilize the lower value of the bound.

$H_{ij}(z)$ contains a pole with multiplicity γ:

Let this pole of multiplicity γ be P. Then, $H_{ij}(z)$ may be expressed as

$$H_{ij}(z) = K_{ij} + \frac{r_{ij}^{(1)}}{(1 - Pz^{-1})} + \frac{r_{ij}^{(2)}}{(1 - Pz^{-1})^2} + \ldots + \frac{r_{ij}^{(\gamma)}}{(1 - Pz^{-1})^{\gamma}},$$

This analysis differs from the one given above for the general term

$$\frac{r_{ij}^{(\zeta)}}{(1 - Pz^{-1})^{\zeta}}$$

where $\zeta = 2, 3, \ldots, \gamma$.

At this point, due mainly to its ease of implementation, we utilize the technique in [10] where the above expression is interpreted as a cascade of ζ first-order sections. For each first-order section, the inverse z-transform is taken using the theory outlined in the distinct pole case. Consider

$$\frac{r_{ij}^{(\zeta)}}{(1 - Pz^{-1})^{\zeta}} = \frac{r_{ij}^{(\zeta)}}{(1 - Pz^{-1})(1 - Pz^{-1}) \ldots (1 - Pz^{-1})}.$$

Taking the inverse z-transform, we get

$$\frac{r_{ij}^{(\zeta)}}{(1 - Pz^{-1})(1 - Pz^{-1}) \ldots (1 - Pz^{-1})} = r_{ij}^{(\zeta)} \cdot \left[\sum_{k=0}^{\infty} |P|^k\right]^{\zeta} = r_{ij}^{(\zeta)} \cdot \left[\frac{1}{1 - |P|}\right]^{\zeta}.$$

This expression is now substituted for the pole of multiplicity γ.

Lemma 1: The zero input response of the state $\mathbf{x}(k)$ of the digital filter described by eqn (7) or (8) is periodic. Its period T satisfies

$$T \leq \prod_{i=1}^{m}(2 \cdot \hat{M}_i + 1) = T_{max}, \tag{16}$$

where $\hat{M}_i$ is the largest integer not more than M_i in eqn (14).

Proof: Consider eqn (7) or (8). The steady-state solution of each state $x_i(k)$ will satisfy

$$|x_i(k)| \leq M_i, \quad \forall k,\ i = 1, 2 \ldots, m.$$

Under fixed-point arithmetic, $\mathbf{x}(k) \in \mathcal{Z}^m$, and hence,

$$|x_i(k)| \leq \hat{M}_i,\ \forall k,\ i = 1, 2 \ldots, m.$$

$x_i(k)$ can therefore take only a finite number of values, namely, $(2 \cdot \hat{M}_i + 1)$. As a result of this, $\mathbf{x}(k)$ can take only a finite number of values, namely,

$$\prod_{i=1}^{m}(2 \cdot \hat{M}_i + 1).$$

Note that, the current state vector $\mathbf{x}(k)$ uniquely determines the next state vector $\mathbf{x}(k+1)$ through the function $\mathcal{Q}[\cdot]$. Thus, $\mathbf{x}(k)$ must be periodic in k. Its period is in fact bounded by

$$T_{max} = \prod_{i=1}^{m}(2 \cdot \hat{M}_i + 1). \tag{17}$$

■

We now have bounds on the amplitude as well as the period on the possible limit cycles. This information will be invaluable for developing our search algorithm.

IV Algorithm Description and Its Computational Aspects

In this section, we formulate the theoretical basis for the algorithm and discuss some of its computational aspects.

Definition 1: The digital filter realization in (9) is said to be globally asymptotically stable (g.a.s.) if and only if, for any initial state $\mathbf{x}(0) \in \mathcal{Z}^m$ with $\|\mathbf{x}(0)\|_\infty \leq B$, where $B \in \mathcal{Z}_+$, there exists $L \in \mathcal{Z}_+$ such that $\mathbf{x}(k) = \mathbf{0}$ for $k \geq L$.

Remark. Typically, g.a.s. is taken to hold when $\mathbf{x}(k) \to \mathbf{0}$ as $k \to \infty$ (under the conditions above). However, due to the finite wordlength available in each register, the digital filter behaves as a finite state machine, and Definition 1 suffices.

Lemma 2: Consider $\eta > 0$ and any initial state vector $\mathbf{x}(0)$ such that

$$|x_i(0)| \leq B_i, \quad \text{for} \quad i = 1, 2, \ldots, m,$$

with $B_i > \hat{M}_i$, for $i = 1, 2, \ldots, m$. Then, there exists a sufficiently large positive number $\mathcal{L}$ such that the digital filter in (7) or (8) satisfies

$$|x_i(k)| \leq \hat{M}_i + \eta, \ \forall k \geq \mathcal{L},$$

for $i = 1, 2, \ldots, m$.

Proof: Since the eigenvalues of A are assumed to lie inside the unit circle in the complex plane, the digital filter in eqn. (9) is in fact g.a.s. Hence, eqn. (9) will yield a set of nonhomogeneous linear shift-invariant difference equations which will have its solution in two parts: A steady-state solution $\mathbf{s}(k)$ and a transient solution $\mathbf{t}(k)$. Clearly, with g.a.s., given $\eta > 0$, we can choose k sufficiently large, say, $k \geq \mathcal{L}$, such that

$$\max |t_i(k)| < \eta, \quad \text{for} \quad i = 1, 2, \ldots, m.$$

Since $\hat{M}_i \in \mathcal{Z}_+$, for $k \geq \mathcal{L}$, $\hat{M}_i + \eta$ will therefore act as a true upper bound for $x_i(k)$ in eqn. (9). ■

Hence, it suffices to check the state vectors in the set $\mathcal{S}^{(0)}$, where

$$\mathcal{S}^{(0)} = \left\{ \mathbf{x}(k) \in \mathcal{Z}^m \,\middle|\, |x_i(k)| \leq \hat{M}_i, \;\; i = 1, 2, \ldots, m \right\}, \tag{18}$$

to see if they are mapped to the zero vector by eqn. (9) after a finite number of mappings.

Computational Aspects

The computations within the algorithm are carried out in two stages. Initially, all vectors $\mathbf{x}(k) \in \mathcal{S}^{(0)}$ which map to $\mathbf{0}$ in less than T_{max} recursions—(after all, if limit cycles exist, the maximum period is T_{max})—are eliminated from $\mathcal{S}^{(0)}$ as they are now known to be stable. The remaining vectors in $\mathcal{S}^{(0)}$ are then further checked for convergence (see Section B).

Section A. Consider the set $\mathcal{V}^{(1)}$, where

$$\mathcal{V}^{(1)} = \left\{ \mathbf{x}(k) \in \mathcal{S}^{(0)} | \mathcal{Q}[A \cdot \mathbf{x}(k)] = \mathbf{0} \right\}, \tag{19}$$

Hence, $\mathcal{V}^{(1)}$ consists of all the vectors $\mathbf{x}(k) \in \mathcal{S}^{(0)}$ that map to $\mathbf{0}$ in one and only one iteration of equation (7) or (8). Note that, any other stable vector in $\mathcal{S}^{(0)}$ must map to $\mathcal{V}^{(1)}$ prior to reaching $\mathbf{0}$. Hence, for further computations, we form

$$\mathcal{S}^{(1)} = \mathcal{S}^{(0)} \setminus \mathcal{V}^{(1)}. \tag{20}$$

Note that, $\mathcal{K}[\mathcal{S}^{(1)}] = \mathcal{K}[\mathcal{S}^{(0)}] - \mathcal{K}[\mathcal{V}^{(1)}]$. In fact, one immediately notices that $\mathcal{K}[\mathcal{S}^{(0)}] = T_{max}$.

Furthermore, any vector in $\mathcal{S}^{(1)}$ which is mapped to $\mathcal{V}^{(1)}$ by (7) or (8) in one iteration will also converge to $\mathbf{0}$. Hence, we form the set $\mathcal{V}^{(2)}$, where

$$\mathcal{V}^{(2)} = \left\{ \mathbf{x}(k) \in \mathcal{S}^{(1)} | \mathcal{Q}[A \cdot \mathbf{x}(k)] \in \mathcal{V}^{(1)} \right\}. \tag{21}$$

Hence, $\mathcal{V}^{(2)}$ consists of all the vectors $\mathbf{x}(k) \in \mathcal{S}^{(1)}$ that map to $\mathbf{0}$ in exactly two iterations of equation (7) or (8). Hence, for further computations, we form

$$\mathcal{S}^{(2)} = \mathcal{S}^{(1)} \setminus \mathcal{V}^{(2)}. \tag{22}$$

Note that, $\mathcal{K}[\mathcal{S}^{(2)}] = \mathcal{K}[\mathcal{S}^{(0)}] - \mathcal{K}[\mathcal{V}^{(1)}] - \mathcal{K}[\mathcal{V}^{(2)}]$.

Likewise, we get the following sets: For $L = 1, 2, \ldots, T_{max}$,

$$\mathcal{V}^{(L)} = \left\{ \mathbf{x}(k) \in \mathcal{S}^{(L-1)} | Q[A \cdot \mathbf{x}(k)] \in \mathcal{V}^{(L-1)} \right\}, \tag{23}$$

and

$$\mathcal{S}^{(L)} = \mathcal{S}^{(L-1)} \setminus \mathcal{V}^{(L)}. \tag{24}$$

Note that, $\mathcal{K}[\mathcal{S}^{(L)}] = \mathcal{K}[\mathcal{S}^{(0)}] - \sum_{i=1}^{L} \mathcal{K}[\mathcal{V}^{(i)}]$.

The conditions under which this construction is terminated and their implications are as follows:
(1) If

$$\mathcal{K}[\mathcal{S}^{(L)}] = \emptyset, \quad \text{for some} \quad L = 1, 2, \ldots, T_{max} - 1, \tag{25}$$

all vectors in $\mathcal{S}^{(0)}$ are convergent.

(2) If

$$\mathcal{K}[\mathcal{V}^{(L)}] = \emptyset, \quad \text{for some} \quad L = 1, 2, \ldots T_{max}, \tag{26}$$

then

$$\mathcal{S}^{(i)} = \mathcal{S}^{(L-1)}, \quad \text{for} \quad i = L, L+1, \ldots, T_{\max}. \tag{27}$$

Under this situation, the remaining vectors in $\mathcal{S}^{(L-1)}$–there are $\mathcal{K}[\mathcal{S}^{(L-1)}]$ of them–will be further checked for convergence (see Section B).

Remark. Upon a little reflection, one notices that $\mathcal{V}^{(T_{max})}$ must either be empty or contain one and only one vector from $\mathcal{S}^{(0)}$.

Section B. Although the reverse mapping procedure outline above reduces the computational complexity considerably, it may not capture all the vectors in $\mathcal{V}^{(L)}$, $L = 1, 2, \ldots, T_{max}$, that map to $\mathbf{0}$ within T_{max} iterations. This is due to the fact that, there may be vectors in $\mathcal{V}^{(L)}$ that map to $\mathbf{0}$ through a vector not belonging to $\mathcal{S}^{(0)}$! Hence, when encountered with condition (2) above, convergence of each remaining vector in $\mathcal{S}^{(L-1)}$ is determined by checking whether it is

mapped to **0** in less than T_{max} through either (7) or (8), whichever is applicable. This exhaustive technique is in fact an extension of that given in [10] to digital filters represented in their state-space realization. However, we must emphasize the significant computational advantage gained by first invoking the reverse mapping construction procedure in Section A.

Assuming condition (2) has occurred, let

$$\mathcal{S}^{(L)} = \left\{ \mathbf{x}_i^{(L)};\ i = 1, 2, \ldots, \mathcal{K}[\mathcal{S}^{(L)}] \right\}. \tag{28}$$

Note that, when condition (2) has occurred, from (27), $\mathcal{S}^{(L-1)} = \mathcal{S}^{(L)}$. For each vector $\mathbf{x}_i^{(L)} \in \mathcal{S}^{(L)}$, construct the orbit $\mathcal{O}_i^{(L)}$ consisting of all state vectors $\mathbf{x}_i^{(L)}(j)$, for $j = 1, 2, \ldots, T_{max}$, that are consecutively generated by (7) or (8) (whichever is applicable) with $\mathbf{x}_i^{(L)}$ as the initial state, that is, $\mathbf{x}_i^{(L)} = \mathbf{x}_i^{(L)}(0)$.

For each $i = 1, 2, \ldots, \mathcal{K}[\mathcal{S}^{(L)}]$, the conditions under which the construction of each orbit $\mathcal{O}_i^{(L)}$ is terminated and their implications are as follows:

(1) If

$$\mathbf{x}_i^{(L)}(j) = 0, \quad \text{for some} \quad j = 1, 2, \ldots, T_{max}, \tag{29}$$

then $\mathbf{x}_i^{(L)}$ together with each vector in the orbit $\mathcal{O}_i^{(L)}$ is convergent.

(2) If

$$\mathbf{x}_i^{(L)}(j) = \mathbf{x}_i^{(L)}(k), \quad \text{for} \quad j \neq k, \tag{30}$$

then $\mathbf{x}_i^{(L)}$ gives rise to limit cycles.

Remark. These are in fact the only conditions that can occur when either (7) or (8) generate the orbit.

Observation.

If the upper bound $\hat{M}_i < 1$ for all i, we observe that $\mathcal{S}^{(0)}$ given by (18) will only contain **0**. Consider a digital filter implementation given by (7), *i.e* $N = 1$. From (14)

$$\varrho \cdot \sum_{j=1}^{m} \sum_{k=0}^{\infty} |h_{ij}(k)| = \hat{M}_i < 1 \tag{31}$$

for $i = 1, 2, \ldots m$,

If a sign magnitude quantizing scheme is considered, $\varrho = 0.5$, equation (31) can be written as,

$$\sum_{j=1}^{m} \sum_{k=0}^{\infty} |h_{ij}(k)| < 2 \tag{32}$$

Therefore we can conclude that a digital filter in double length accumulator environment satisfying eqn. (32), is globally asymptotically stable.

V Perturbation of Filter Coefficient Matrix

In constructing the region of g.a.s. in the coefficient space, perturbations incurred in storing each filter coefficient must also be considered. Such perturbations are typically due to finite wordlength effects that require rounding or truncation of the true coefficient value.

The algorithm described in the previous section provides information regarding g.a.s. of a given filter with a nominal coefficient matrix $A = \{a_{ij}\} \in \Re^{m \times m}$. Once this is done, we now consider a small perturbation Δa_{ij} of each coefficient about its nominal value a_{ij}. However, for a given state vector $\mathbf{x}(k)$, this perturbation may not necessarily alter the next state $\mathbf{x}(k+1)$ obtained since it is entirely possible that

$$\mathbf{x}(k+1) = \mathcal{Q}[(A + \Delta A) \cdot \mathbf{x}(k)] = \mathcal{Q}[A \cdot \mathbf{x}(k)], \tag{33}$$

where $\Delta A = \{\Delta a_{ij}\} \in \Re^{m \times m}$.

Depending on the number of quantizers per row, that is, depending on whether a double- or single-length accumulator is available, (33) is interpreted differently.

Double-length accumulator

It is evident that the upper bound $\hat{M}_i$ estimated for the nominal value of the coefficient matrix $\{a_{ij}\} \in \Re^{m \times m}$ will no longer be valid for a perturbed system $\{a_{ij} + \Delta a_{ij}\} \in \Re^{m \times m}$. We

define an upper bound $\tilde{M}_i$, $i = 1, 2, \ldots, m$, which will be valid for the nominal coefficient matrix $\{a_{ij}\}_{m\times m}$ and a perturbed coefficient matrix $\{a_{ij} + \Delta a_{ij}\}_{m\times m}$. Consider the equation,

$$\mathcal{Q}\left[\sum_{j=1}^{m}(a_{ij} + \Delta a_{ij}) \cdot x_j(k)\right] = \mathcal{Q}\left[\sum_{j=1}^{m} a_{ij} \cdot x_j(k)\right], \tag{34}$$

$i = 1, 2, \ldots, m$ and where $x_j \in \mathbf{x}$ is taken from the set,

$$\tilde{\mathcal{S}} = \left\{\mathbf{x} \,\middle|\, |x_i| \leq \tilde{M}_i, \;\; i = 1, 2, \ldots, m; \; \mathbf{x} \in \mathcal{Z}^m\right\}. \tag{35}$$

Due to the choice of $\tilde{M}_i$, $\tilde{\mathcal{S}}$ is valid for the systems described by (7) with coefficient matrices $\{a_{ij}\}_{m\times m}$ and $\{a_{ij} + \Delta a_{ij}\}_{m\times m}$. Let $\mathcal{G}$ be the set consisting of all the perturbations Δa_{ij} around the nominal value a_{ij}, which satisfies equation (34). This is in fact the *Robustness region* associated with the nominal value of the coefficient matrix $\{a_{ij}\}$. We formulate the problem of finding the robustness region in the following manner.

Any perturbation Δa_{ij} satisfying eqn. (34) for all $\mathbf{x} \in \tilde{\mathcal{S}}$ will be in the set $\mathcal{G}$.

Since it is not possible to consider an arbitrarily large area around the nominal coefficient value, we use an analytical method to make an estimate for this region to which the algorithm can be applied. Once the region is determined it will be covered by a suitable grid and eqn. (34) will be used to determine if each grid point, corresponding to a particular Δa_{ij} in this estimated region, is in fact in $\mathcal{G}$.

To proceed, it is convenient to identify the discontinuities associated with the nonlinearity $\mathcal{Q}[\cdot]$.

For sign-magnitude roundoff,

$$\mathcal{D}_r = \left\{b_r \in \Re \mid b_r = r + \frac{1}{2}, \; r \in \mathcal{Z}\right\}; \tag{36}$$

for sign-magnitude truncation quantization,

$$\mathcal{D}_{mt} = \{b_r \in \Re \mid b_r = r, \; r \in \mathcal{Z} \setminus \{0\}\}; \tag{37}$$

for two's complement truncation quantization,

$$\mathcal{D}_{two} = \{b_r \in \Re \mid b_r = r, \; r \in \mathcal{Z}\}. \tag{38}$$

For each $\mathbf{x} \in \tilde{\mathcal{S}}$, a region $\mathcal{G}_{\mathbf{x}}$ corresponding to the robustness region in eqn. (34) applicable to the pertinent quantization schemes in (36), (37), or (38) is defined. Let the region corresponding to the i-th state x_i of $\mathbf{x}$ be $\mathcal{G}_{\mathbf{x}}^{(i)}$. Then, we have the following:

For sign-magnitude roundoff quantization,

$$\mathcal{G}_{\mathbf{x}}^{(i)} = \left\{ \begin{array}{ll} \{\Delta a_{ij} | & b_{r-1} - \sum_{j=1}^{m} a_{ij}x_j \leq \sum_{j=1}^{m} \Delta a_{ij}x_j < b_r - \sum_{j=1}^{m} a_{ij}x_j\} \\ & \text{for } b_{r-1} \leq \sum_{j=1}^{m} a_{ij}x_j < b_r \text{ and } r \geq 1 \\ \{\Delta a_{ij} | & b_{r-1} - \sum_{j=1}^{m} a_{ij}x_j < \sum_{j=1}^{m} \Delta a_{ij}x_j \leq b_r - \sum_{j=1}^{m} a_{ij}x_j\} \\ & \text{for } b_{r-1} < \sum_{j=1}^{m} a_{ij}x_j \leq b_r \text{ and } r \leq -1 \\ \{\Delta a_{ij} | & b_{-1} - \sum_{j=1}^{m} a_{ij}x_j < \sum_{j=1}^{m} \Delta a_{ij}x_j < b_0 - \sum_{j=1}^{m} a_{ij}x_j\} \\ & \text{for } b_{-1} < \sum_{j=1}^{m} a_{ij}x_j < b_0 \end{array} \right\}$$

$$\text{where} \quad b_r \in \mathcal{D}_r; \tag{39}$$

for sign-magnitude truncation quantization,

$$\mathcal{G}_{\mathbf{x}}^{(i)} = \left\{ \begin{array}{ll} \{\Delta a_{ij} | & b_{r-1} - \sum_{j=1}^{m} a_{ij}x_j \leq \sum_{j=1}^{m} \Delta a_{ij}x_j < b_r - \sum_{j=1}^{m} a_{ij}x_j\} \\ & \text{for } b_{r-1} \leq \sum_{j=1}^{m} a_{ij}x_j < b_r \text{ and } r \geq 2 \\ \{\Delta a_{ij} | & b_{r-1} - \sum_{j=1}^{m} a_{ij}x_j < \sum_{j=1}^{m} \Delta a_{ij}x_j \leq b_r - \sum_{j=1}^{m} a_{ij}x_j\} \\ & \text{for } b_{r-1} < \sum_{j=1}^{m} a_{ij}x_j \leq b_r \text{ and } r \leq -1 \\ \{\Delta a_{ij} | & b_{-1} - \sum_{j=1}^{m} a_{ij}x_j < \sum_{j=1}^{m} \Delta a_{ij}x_j < b_{+1} - \sum_{j=1}^{m} a_{ij}x_j\} \\ & \text{for } b_{-1} < \sum_{j=1}^{m} a_{ij}x_j < b_{+1} \end{array} \right\}$$

$$\text{where} \quad b_r \in \mathcal{D}_{mt}; \tag{40}$$

for two's complement truncation quantization,

$$\mathcal{G}_{\mathbf{x}}^{(i)} = \left\{ \begin{array}{ll} \{\Delta a_{ij} | & b_{r-1} - \sum_{j=1}^{m} a_{ij}x_j \leq \sum_{j=1}^{m} \Delta a_{ij}x_j < b_r - \sum_{j=1}^{m} a_{ij}x_j\} \\ & \text{for } b_{r-1} \leq \sum_{j=1}^{m} a_{ij}x_j < b_r \end{array} \right\}$$

$$\text{where} \quad b_r \in \mathcal{D}_{two}. \tag{41}$$

For a particular state in the vector $\mathbf{x}$, the region can be computed using eqns.(39) , (40) or (41). For all $\mathbf{x} \in \tilde{\mathcal{S}}$ the total robustness region is given by

$$\mathcal{G} = \bigcap_{\mathbf{x} \in \tilde{\mathcal{S}}} \mathcal{G}_{\mathbf{x}}. \tag{42}$$

From (39) , (40) or (41), we may estimate suitable values for the region of robustness for each quantization scheme. The computations involved in determining the robustness region for the two's complement quantization scheme is given below, the analysis can be extended for the other quantization schemes in a similar manner. For the two's complement truncation quantization scheme from eqn.(41) , we choose Δa_{ij} such that,

$$\sum_{j=1}^{m} \Delta a_{ij} x_j \leq \min_{\mathbf{x} \in \tilde{\mathcal{S}}} \left\{ |b_r - \sum_{j=1}^{m} a_{ij} x_j|, |b_{r-1} - \sum_{j=1}^{m} a_{ij} x_j| \right\} \tag{43}$$

$i = 1, 2, \ldots, m$. The left hand side of (43) will be given by,

$$\left| \sum_{j=1}^{m} \Delta a_{ij} x_j \right| \leq \sum_{j=1}^{m} |\Delta a_{ij}| \cdot |x_j| \tag{44}$$

We estimate the perturbations Δa_{ij} such that, they satisfy the following equation,

$$\sum_{j=1}^{m} |\Delta a_{ij}| \cdot |x_j| \leq \min_{\mathbf{x} \in \tilde{\mathcal{S}}} \left\{ |b_r - \sum_{j=1}^{m} a_{ij} x_j|, |b_{r-1} - \sum_{j=1}^{m} a_{ij} x_j| \right\} \tag{45}$$

where $i = 1, 2, \ldots, m$.

Since we are estimating the region for the nominal value in the coefficient space we will initially take each x_j to be bounded by its corresponding $\hat{M}_j$. Therefore the above equation is satisfied if.

$$\sum_{j=1}^{m} |\Delta a_{ij}| \cdot |\hat{M}_j| \leq \min_{\mathbf{x} \in \tilde{\mathcal{S}}} \left\{ |b_r - \sum_{j=1}^{m} a_{ij} x_j|, |b_{r-1} - \sum_{j=1}^{m} a_{ij} x_j| \right\} \tag{46}$$

Now (46) can be used to estimate the robustness region $\hat{\mathcal{G}}$,where $\hat{\mathcal{G}}$ is given by,

$$\hat{\mathcal{G}} = \left\{ \Delta a_{ij} \left| \sum_{j=1}^{m} |\Delta a_{ij}| \cdot |\hat{M}_j| \leq \min_{\mathbf{x} \in \tilde{\mathcal{S}}} \{ |b_r - \sum_{j=1}^{m} a_{ij} x_j|, |b_{r-1} - \sum_{j=1}^{m} a_{ij} x_j| \} \right. \right\} \tag{47}$$

Where $i = 1, 2, \ldots, m$. Clearly, $\hat{\mathcal{G}} \subset \mathcal{G}$.

But from eqn.(47) it is observed that in a degenerate case $\hat{\mathcal{G}}$ may only contain the zero perturbation vector, due to the right hand side of eqn (47) being equal to zero.

Note that for all quantization schemes considered, and for all i,

$$\min_{\mathbf{x} \in \tilde{\mathcal{S}}} \left\{ |b_r - \sum_{j=1}^{m} a_{ij} x_j|, |b_{r-1} - \sum_{j=1}^{m} a_{ij} x_j| \right\} < \frac{1}{2} \tag{48}$$

due to the distance between any two discontinuities being always less than 1. Therefore the maximum perturbation region, $\tilde{\mathcal{G}}$ is given by the following set and it is seen that $\hat{\mathcal{G}} \subset \tilde{\mathcal{G}}$,

$$\tilde{\mathcal{G}} = \left\{ \Delta a_{ij} \left| \sum_{j=1}^{m} |\Delta a_{ij}||\hat{M}_j| < \frac{1}{2} \right. \right\} \tag{49}$$

Where $\hat{M}_j$ for $j = 1, 2 \ldots, m$ are the upper bounds computed for the nominal value $\{a_{ij}\}$.

Single-length accumulator

If there are m quantizers per row as in (8), robustness region is defined for each element a_{ij} in the following manner:

$$\mathcal{G} = \left\{ \Delta a_{ij} | \mathcal{Q}\left[(a_{ij} + \Delta a_{ij}) \cdot x_j\right] = \mathcal{Q}\left[(a_{ij} \cdot x_j)\right], \quad \forall \mathbf{x} \in \tilde{\mathcal{S}}_1 \right\}. \tag{50}$$

$i, j = 1, 2, \ldots, m$

As in the double length accumulator implementation we define an upper bound $\tilde{M}_i$, valid for systems with coefficient matrices $\{a_{ij}\}$ and $\{a_{ij} + \Delta a_{ij}\}$ and described by eqn. (8). The set $\tilde{\mathcal{S}}_1$ is defined as follows,

$$\tilde{\mathcal{S}}_1 = \left\{ \mathbf{x} \left| |x_i| \leq \tilde{M}_i \; ; \quad i = 1, 2, \ldots, m; \quad \mathbf{x} \in \mathcal{Z}^m \right. \right\} \tag{51}$$

Let the robustness region corresponding to element a_{ij} in the coefficient matrix for a particular state vector $\mathbf{x}$ be $\mathcal{G}_{\mathbf{x}}^{(i,j)}$. Then, we have the following:

For sign-magnitude roundoff quantization,

$$\mathcal{G}_{\mathbf{x}}^{(i,j)} = \left\{ \begin{array}{ll} \{\Delta a_{ij} | & b_{r-1} - a_{ij} \cdot x_j \leq \Delta a_{ij} \cdot x_j < b_r - a_{ij} \cdot x_j\} \\ & \text{for } b_{r-1} \leq a_{ij} \cdot x_j < b_r \text{ and } r \geq 1 \\ \{\Delta a_{ij} | & b_{r-1} - a_{ij} \cdot x_j < \Delta a_{ij} \cdot x_j \leq b_r - a_{ij} \cdot x_j\} \\ & \text{for } b_{r-1} < a_{ij} \cdot x_j \leq b_r \text{ and } r \leq -1 \\ \{\Delta a_{ij} | & b_{-1} - a_{ij} \cdot x_j < \Delta a_{ij} \cdot x_j < b_0 - a_{ij} \cdot x_j\} \\ & \text{for } b_{-1} < a_{ij} \cdot x_j < b_0 \end{array} \right\}$$

$$\text{where} \quad b_r \in \mathcal{D}_r, \quad \forall \mathbf{x} \in \tilde{\mathcal{S}}_1; \tag{52}$$

for sign-magnitude truncation quantization,

$$\mathcal{G}_{\mathbf{x}}^{(i,j)} = \left\{ \begin{array}{ll} \{\Delta a_{ij} | & b_{r-1} - a_{ij} \cdot x_j \leq \Delta a_{ij} \cdot x_j < b_r - a_{ij} \cdot x_j\} \\ & \text{for } b_{r-1} \leq a_{ij} \cdot x_j < b_r \text{ and } r \geq 2 \\ \{\Delta a_{ij} | & b_{r-1} - a_{ij} \cdot x_j < \Delta a_{ij} \cdot x_j \leq b_r - a_{ij} \cdot x_j\} \\ & \text{for } b_{r-1} < a_{ij} \cdot x_j \leq b_r \text{ and } r \leq -1 \\ \{\Delta a_{ij} | & b_{-1} - a_{ij} \cdot x_j < \Delta a_{ij} \cdot x_j < b_{+1} - a_{ij} \cdot x_j\} \\ & \text{for } b_{-1} < a_{ij} \cdot x_j < b_{+1} \end{array} \right\}$$

$$\text{where} \quad b_r \in \mathcal{D}_{mt} \quad \forall \mathbf{x} \in \tilde{\mathcal{S}}_1 \tag{53}$$

For two's complement truncation quantization,

$$\mathcal{G}_{\mathbf{x}}^{(i,j)} = \left\{ \begin{array}{ll} \{\Delta a_{ij} | & b_{r-1} - a_{ij} \cdot x_j \leq \Delta a_{ij} \cdot x_j < b_r - a_{ij} \cdot x_j\} \\ & \text{for } b_{r-1} \leq a_{ij} \cdot x_j < b_r \end{array} \right\}$$

$$\text{where} \quad b_r \in \mathcal{D}_{two} \quad \forall \mathbf{x} \in \tilde{\mathcal{S}}_1. \tag{54}$$

Hence g.a.s. can be gaurenteed for the region

$$\mathcal{G} = \bigcap_{\forall (i,j)} \mathcal{G}_{\mathbf{x}}^{(i,j)}. \tag{55}$$

Using a similar argument as in the case of a double-length accumulator, we can estimate a region of robustness for each quantization scheme using eqns. (52), (53) or (54). The computations involved for the two's complement case is outlined below. The perturbation is seen to satisfy the equation,

$$\Delta a_{ij} x_j \leq \min_{\mathbf{x} \in \tilde{\mathcal{S}}_1} \{|b_r - a_{ij} x_j|, |b_{r-1} - a_{ij} x_j|\}, \quad \text{for all } i, j \tag{56}$$

$|\Delta a_{ij} x_j| = |\Delta a_{ij}| \cdot |x_j|$ and the right hand side of eqn. (56) satisfies (48) therefore it can be rewritten in the following form,

$$|\Delta a_{ij}| \cdot |x_j| \leq \frac{1}{2} \tag{57}$$

Since we are only interested in finding an estimate for the region around the nominal value of the coefficient, any x_j is bounded by it's corresponding $\hat{M}_i$. The estimated region for the two's complement quantization will consist of any perturbation satisfying th equation,

$$\tilde{\mathcal{G}} = \left\{ \Delta a_{ij} \left| |\Delta a_{ij}||\hat{M}_i| < \frac{1}{2}\,; \quad i, j = 1, 2, \ldots, m \right. \right\} \tag{58}$$

The region $\tilde{\mathcal{G}}$ can be covered by a suitable grid and the grid points are applied to eqn. (50) to obtain the stable region around the nominal point.

We note that if $\hat{M}_i < 1$ for all i then the perturbations Δa_{ij} $(i, j = 1, 2, \ldots, m)$ can take any value.

VI Some Examples

In this section the proposed search algorithm is applied to a dense grid in the coefficient space to obtain the total global asymptotic stability region for a digital filter with zero input. The dense grid will provide a reasonably good approximation to the g.a.s region, since it is not possible to consider all points in the linear stability region. Note that each point in the coefficient space is associated with a neighborhood where the filter is stable. A 10 Bit wordlength is assume for all computations, therefore the filter coefficients are quantized to a multiple of 2^{-10}. Within the linear stability region dark areas indicate points where limit cycles of some period exists. It should be noted that the linear stability region does not have a common boundary with the global asymptotic stability region obtained through this algorithm. Therefore in all figures, the boundary line which delimits the stability region from the unstable region does not belong to the stability region.

The most commonly encountered quantization schemes are analyzed, they are namely, sign magnitude roundoff quantization scheme, sign-magnitude truncation quantization scheme and the two's complement truncation quantization scheme. In all quantization schemes the single- and the double-length accumulator implementation results are provided. All results are provided for the $\{a_{ij}\} \in \Re^{2\times 2}$ coefficient matrix. All existing results for the named quantization schemes were verified.

Results for direct form realization of digital filters

For a direct form digital filter in state space formulation (the coefficient matrix is given by eqn.(59))

$$A = \begin{bmatrix} 0 & 1 \\ a_2 & a_1 \end{bmatrix} \tag{59}$$

Figure.(1) shows the region obtained by the proposed algorithm the sign magnitude roundoff quantization scheme in an double length accumulator environment.

The region obtained is identical to the results given in [10]. For the same quantization scheme and single length accumulator the region obtained is given in Figure.(2). The region matches exactly with the ones found in [10]. The regions for the two's complement and the sign magnitude truncation schemes were also verified. The region for the two's complement truncation quantization in the single length accumulator environment is shown in Figure (3) .Note that there is a graphical error in the region given in [10]. All other regions obtained by the proposed algorithm matches with the regions given in [10].

Results for minimum norm realization of digital filters

The stability of digital filters in its minimum norm form for the coefficient matrix, $A \in \Re^{2\times 2}$ case was also investigated. The coefficient matrix is given by eqn.(60).

$$A = \begin{bmatrix} \sigma & \omega \\ -\omega & \sigma \end{bmatrix} \tag{60}$$

The results for the sign magnitude roundoff scheme for the single- and the double-length accumulator environment is given in Figure.(4) This region matches with the region given in [7]. The stable region for the sign magnitude truncation scheme in a single length or double length accumulator environment spans the entire region where $\sigma^2 + \omega^2 < 1$. results are given in Figure.(5). This supports previously known results.

For the two's complement truncation quantization, with double length accumulator the global asymptotic region is given in Figure.(6). This supports and also improves on the previously

known results given in [19]. To the authors knowledge no previous results are available for the two's complement truncation quantization in a single length accumulator environment. The region of global asymptotic stability is summarized in Figure.(7).

Note that for the Two's complement quantization scheme in a double length accumulator environment, series of points extend from the stability region into the instability region such that,

$$\sigma < 0 \quad \text{and} \quad \omega = \pm\sigma \tag{61}$$

The following coefficient matrix can be cited as an example,

$$A = \begin{bmatrix} -\frac{672}{1024} & \frac{672}{1024} \\ -\frac{672}{1024} & -\frac{672}{1024} \end{bmatrix} \tag{62}$$

This series of points can only be observed by magnifying the area concerned. Sub figures shown in Figure (6) shows these areas magnified.

Robustness regions

Some examples of the robustness regions computed using the theory outlined in Section V is given below.

The robustness region for,

$$A = \begin{bmatrix} 0 & 1 \\ \frac{102}{1024} & \frac{102}{1024} \end{bmatrix} \tag{63}$$

for the sign magnitude roundoff quantization scheme in double- and single- length accumulator environments are given in Figure.(8) and Figure.(9) respectively. The robustness region associated with the coefficient matrix given by eqn. (62) for the two's complement quantization scheme under double length accumulator environment is given in Figure.(10).

VII Conclusion

A new algorithm capable of determining global asymptotic stability of any fixed point digital filter represented in its state space formulation, under zero input conditions has been presented. The search algorithm is independent of the type of nonlinearity, the number of nonlinearities and it has been generalized to handle a digital filter of order m in its state space represented form.

The proposed algorithm is found to provide tighter bounds on the amplitude of limit cycles in most cases, and it will always determine the stability or instability of a particular digital filter. Significant improvement over the existing results for the two's complement truncation schemes in both single- and double length accumulator environments have been presented.

The current research is directed towards the following problems.

(1) Establishing regions within which limit cycles of a pre-specified period exists.

(2) Establish regions within which limit cycles that are under a pre-specified bound exist.

(3) Extension of the algorithm for δ-operator formulated systems. In Fixed-point arithmetic it is known that such systems always exhibit limit cycle behavior [20]. Therefore in actual applications the regions similar to the ones mentioned in items (1) and (2) may be of importance.

References

[1] E.I. Jury and B.W. Lee, "The absolute stability of systems with many nonlinearities," *Automat. Remote Contr.,* vol 26, no. 6, pp. 943-961, 1965.

[2] W. Barnes and A.T. Fam, "Minimum norm recursive digital filters that are free of overflow oscillations," *IEEE Trans. Circ. Syst.,* vol. CAS-24, no. 10, pp. 569-574, Oct. 1977.

[3] W.L. Mills, C.T. Mullis, and R.A. Roberts, "Digital filter realizations without overflow oscillations," *Proc. 1978 IEEE Int. Conf. Acoust., Speech, Sig. Proc.,* pp. 71-74, 1978.

[4] T. Claasen, W.F.G. Mecklenbräuker, and J.B.H. Peek, "Frequency domain criteria for the absence of zero-input limit cycles in nonlinear discrete-time systems with applications to digital filters," *IEEE Trans. Circ. Syst.*, vol. CAS-22, no. 3, pp. 232-239, Mar. 1974.

[5] E.D. Garber, "Frequency criteria for the absence of periodic responses," *Automat. Remote. Contr.*, vol. 28, no. 11, pp. 1776-1780, 1967.

[6] V. Singh, "An extension to Jury-Lee's criterion for the stability analysis of fixed-point digital filters designed with two's complement arithmetic," *IEEE Trans. Circ. Syst.*, vol. CAS-33, no. 3, p. 355, Mar. 1986.

[7] K.T. Erickson and A.N. Michel, "Stability analysis of fixed-point digital filters using computer generated Lyapunov functions—Part I: Direct form and coupled form filters," *IEEE Trans. Circ. Syst.*, vol. CAS-32, no. 2, pp. 113-131, Feb. 1985.

[8] K.T. Erickson and A.N. Michel, "Stability analysis of fixed-point digital filters using computer generated Lyapunov functions—Part II: Wave digital filters and lattice digital filters," *IEEE Trans. Circ. Syst.*, vol. CAS-32, no. 2, pp. 132-142, Feb. 1985.

[9] A.N. Michel and R.K. Miller, "Stability analysis of discrete time interconnected systems via computer generated Lyapunov functions with applications to digital filters," *IEEE Trans. Circ. Syst.*, vol. CAS-32, no. 8, pp. 737-753, Aug. 1985.

[10] P.H. Bauer and L.J. Leclerc, "A computer-aided test for the absence of limit cycles in fixed-point digital filters," *IEEE Trans. Sig. Proc.*, vol. 39. no. 11, pp. 2400-2409, Nov. 1991.

[11] J.F. Kaiser, "Some special practical considerations in the realization of linear digital filters," *Proc. 3rd Allerton Ann. Conf. Circ. Syst. Theory,* pp. 100-104, 1965.

[12] T. Bose and D.P. Brown, "Limit cycles in zero input digital filters due to two's complement quantization," *IEEE Trans. Circ. Syst.*, vol. CAS-37, no. 4, pp. 568-571, Month April 1990.

[13] A. Lepschy, G.A. Mian and U. Viaro, "Effects of quantization in second-order fixed-point digital filters with two's complement truncation quantization," *IEEE Trans. Circ. Syst.*, vol. CAS-35, no. 4, pp. 461-466, April 1988.

[14] Trân-Thông and B. Liu, "Limit cycles in the combination implementation of digital filters," *IEEE Trans. Acoust., Speech, Sig. Proc.*, vol. ASSP-24, no. 3, pp. 248-256, Feb. 1976.

[15] Trân-Thông and B. Liu, "A contribution to the stability analysis of second-order direct-form digital filters with magnitude truncation," *IEEE Trans. Acoust., Speech, Sig. Proc.*, vol. ASSP-35, no. 8, pp. 1207-1210, Aug. 1987.

[16] Trân-Thông and B. Liu, "Parameter space quantization in fixed-point digital filters," *Electron. Lett.*, vol. 22, no. 7, pp. 384-386, Mar. 1986.

[17] Trân-Thông and B. Liu, "Parameter plane quantization induced by signal quantization in second-order fixed-point digital filters with one quantizer," *Sig. Proc.*, vol. 14, no. 1, pp. 103-106, Jan. 1988.

[18] Trân-Thông and B. Liu, "Zero-input limit cycles and stability in second order fixed point digital filters with two magnitude truncation quantizers," *Circ., Syst., Sig. Proc.*, vol. 8, no. 4, 1989.

[19] T. Bose, "Stability of digital filters implemented with Two's complement truncation quantization," *IEEE Trans. Sig. Proc.*, vol. 40. no. 01, pp. 24-31, Jan. 1992.

[20] K. Premaratne, P. H. Bauer, "Limit cycles and asymptotic stability of delta-operator formulator discrete time systems implemented in fixed point arithmetic," *Proc. IEEE intl. Symp. on Circ. Syst. ISCAS '94 London UK*, pp. 461-464 ,May-June 1994.

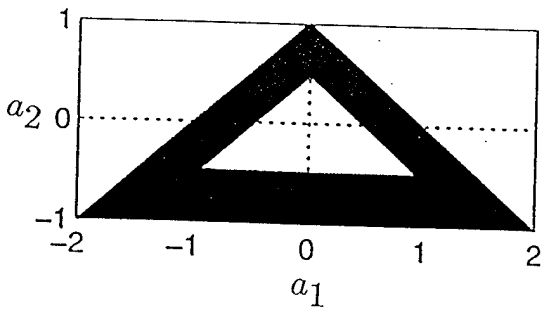


Figure 1: Region where a direct form digital filter is free of limit cycles for the sign magnitude roundoff quantization scheme in a double length accumulator environment.

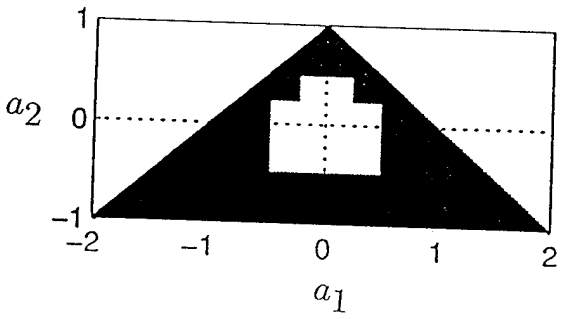


Figure 2: Region where a direct form digital filter is free of limit cycles for the sign magnitude roundoff quantization scheme in a single length accumulator environment.

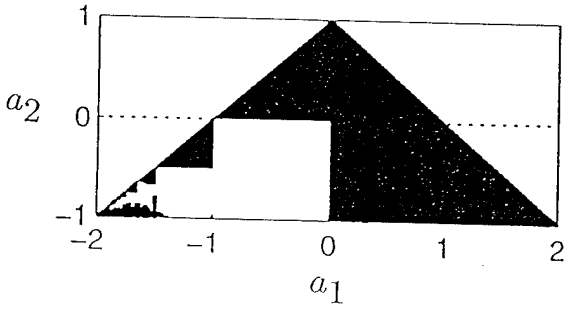


Figure 3: Region where a direct form digital filter is free of limit cycles for the two's complement truncation quantization scheme in a single length accumulator environment.

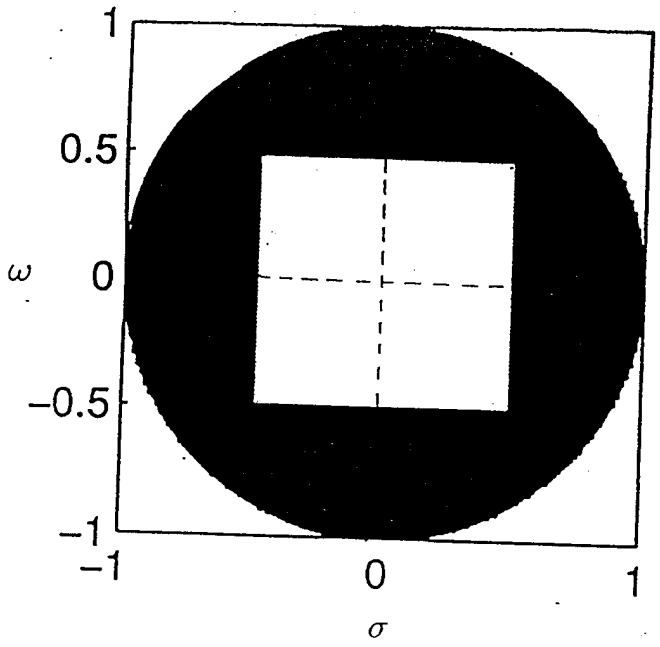


Figure 4: Region where a filter represented in minimum norm form is free of limit cycles for sign magnitude roundoff quantization, in double and single length accumulator environments.

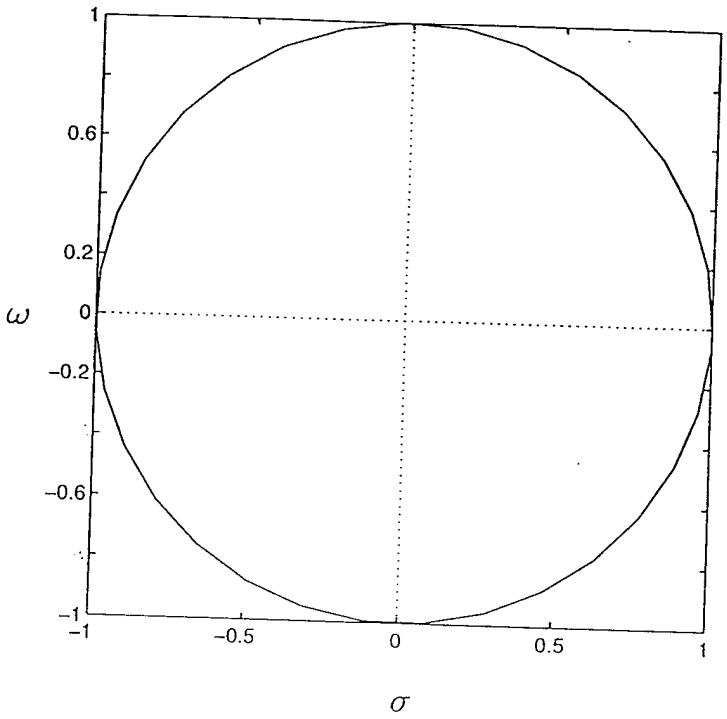
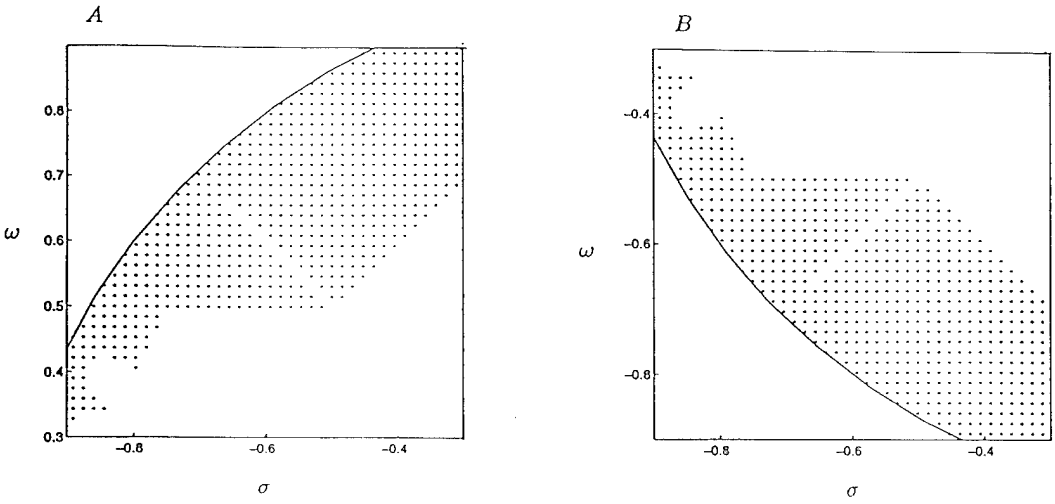


Figure 5: Region where a filter represented in minimum norm form is free of limit cycles for sign magnitude truncation quantization, in double and single length accumulator environments.



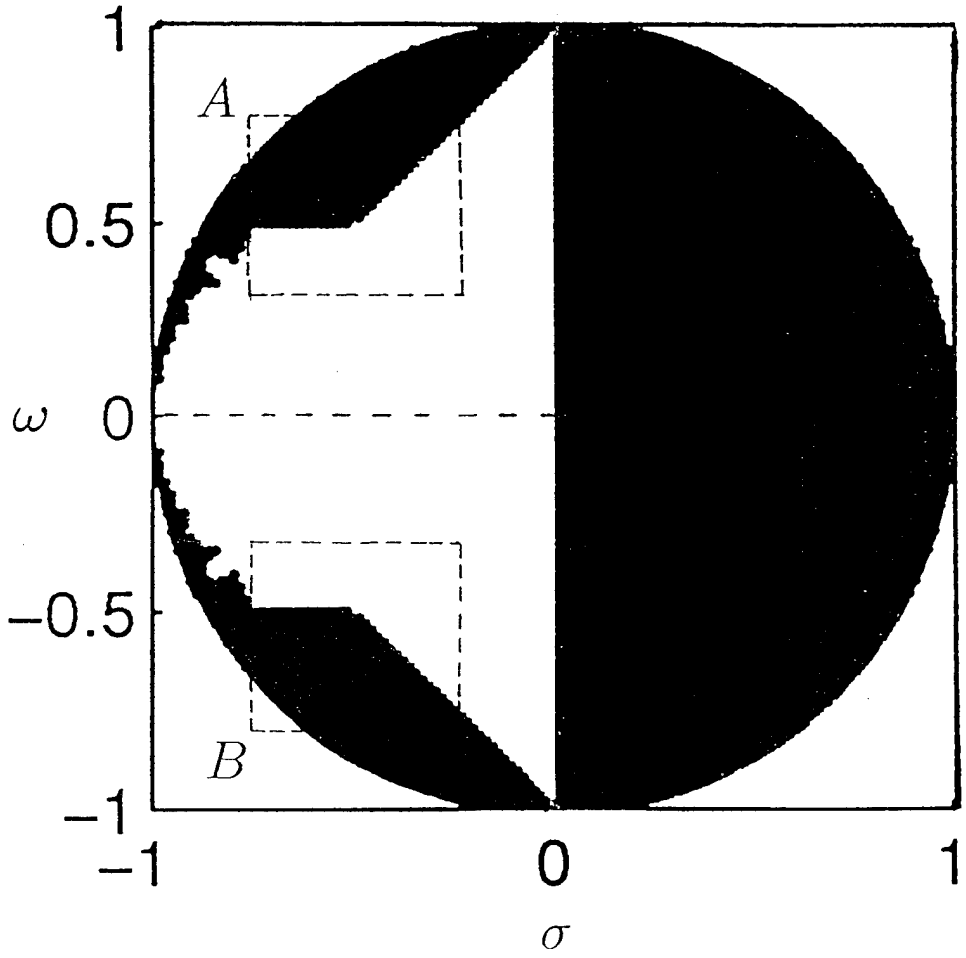


Figure 6: Region where a filter represented in minimum norm form is free of limit cycles for two's complement truncation quantization, in a double length accumulator environment.

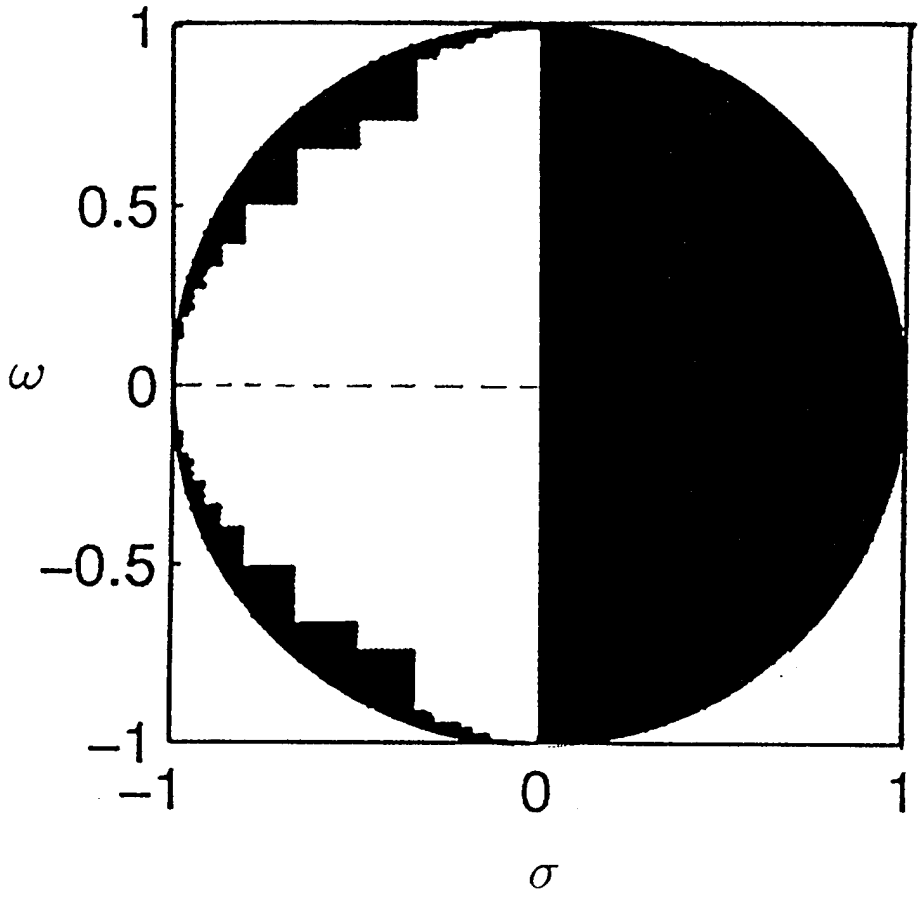


Figure 7: Region where a filter represented in minimum norm form is free of limit cycles for two's complement truncation quantization, in a single length accumulator environment.

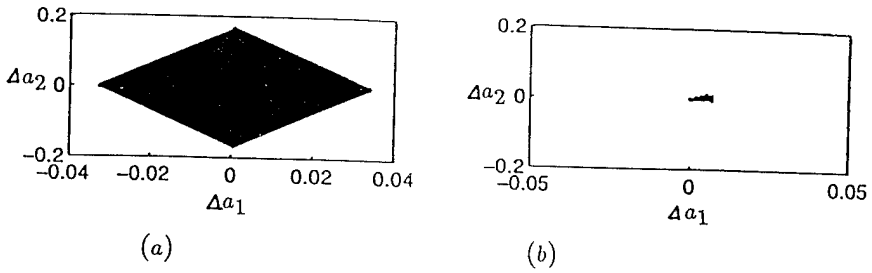


Figure 8: (a) Estimated robustness region (b) Actual robustness region, for a filter analyzed using roundoff quantization in a double length accumulator environment.

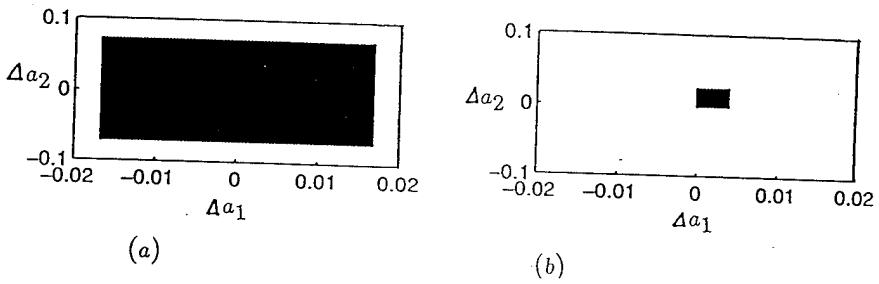


Figure 9: (a) Estimated robustness region (b) Actual robustness region, for a filter analyzed using roundoff quantization in a single length accumulator environment.

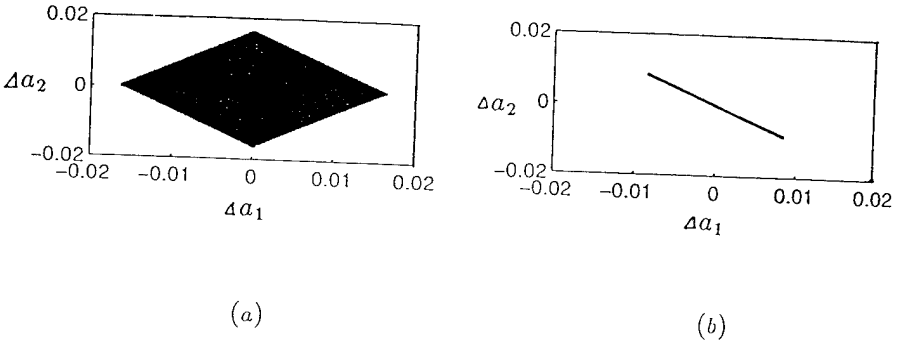


Figure 10: (a) Estimated robustness region. (b) Actual robustness region, for the coefficient matrix given in eqn. (62).

IEEE Transactions on Circuits and Systems---I: Fundamental Theory and Applications, in review, 1995

Limit Cycles in Delta-Operator Formulated 1-D and M-D Discrete-Time Systems with Fixed-Point Arithmetic

Peter H. Bauer
Department of Electrical Engineering
University of Notre Dame
Notre Dame, IN 46556

Kamal Premaratne
Department of Electrical and Computer Engineering
University of Miami
Coral Gables, FL. 33124

ABSTRACT

In this paper, the problem of global asymptotic stability of δ-operator formulated one-dimensional (1-D) and multi-dimensional (m-D) discrete-time systems is analyzed for the case of fixed point implementations. It is shown that the free response of such a system tends to produce incorrect equilibrium points if conventional quantization arithmetic schemes such as truncation or rounding are used. Explicit necessary conditions for global asymptotic stability are derived in terms of the sampling period. These conditions demonstrate that, in almost all cases, fixed-point arithmetic does not allow for global asymptotic stability in δ-operator formulated discrete-time systems that use a short sampling time. This is true for the 1-D as well as the m-D case.

I. INTRODUCTION

Discrete-time systems formulated in terms of the incremental difference operator (or, δ-operator) have recently been receiving considerable attention in the technical literature [1-4]. Most of this work focuses on the superior performance of the δ-operator under finite wordlength conditions when compared with the shift-operator (or, q-operator). In particular, investigations of coefficient sensitivity and quantization noise properties have revealed that δ-operator formulations usually perform significantly better than their q-operator counterparts [1-4]. This is especially true for high-speed applications where the sampling rate is much larger than the underlying system bandwidth. Under these conditions, q-operator formulated discrete-time systems tend to become ill-conditioned [1-2].

Although a large amount of work is available on the effects of coefficient sensitivity and quantization noise, a deterministic study of the nonlinear behavior of discrete-time systems formulated with the δ-operator has not been undertaken. In the case of floating-point (FLP) arithmetic, some results for feedback system are available in [2].

In this work, we focus on the convergence behavior of the unforced system response and global asymptotic stability of δ-operator formulated discrete-time systems implemented in fixed-point (FXP) arithmetic. In particular, via necessary conditions for stability, it will be shown that such systems tend to produce DC limit cycles. We will also perform a deterministic analysis of the finite wordlength properties of multi-dimensional δ-operator implemented discrete time systems. The stability behavior in the m-D case has not been previously investigated, although convergence to the true equilibrium point(s) is one of the most fundamental requirements for any discrete time system realization.

The structure of this article is as follows: In Section II, we introduce notation and nomenclature for the 1-D case. The model for 1-D δ-operator formulated discrete-time systems, with and without quantization nonlinearities, is briefly discussed. Section III addresses the problem of asymptotic stability for the 1-D case. In terms of ensuing DC limit cycles, necessary conditions for global asymptotic stability are formulated. It is shown that, when FXP arithmetic is used, stability of the linear system is often lost. Bounds on the

size of the deadbands are also provided. In section IV, the multidimensional case is investigated using sets of 1-D conditions for asymptotic stability. Section V provides concluding remarks.

II. NOTATION AND NOMENCLATURE

Since our focus is the investigation of stability properties of δ-operator formulated discrete-time systems under unforced conditions, the state equations of the system under zero-input will be considered.

In the linear case, the general m-th order state-space representation is given by

$$\delta[\mathbf{x}](n) = A^{\delta}\mathbf{x}(n); \tag{1}$$

$$\mathbf{x}(n+1) = \mathbf{x}(n) + \Delta \cdot \delta[\mathbf{x}](n), \tag{2}$$

where $\mathbf{x}(n) = [x_1(n), \ldots, x_m(n)]^T$ is the state vector at instant n, $A^{\delta} = \{a_{ij}^{\delta}\} \in \Re^{m\times m}$ is the system matrix, and $\Delta > 0$ is the sampling time. Moreover, $\delta[\cdot]$ represents the δ-operator, that is,

$$\delta[x_\nu](n) = \frac{x_\nu(n+1) - x_\nu(n)}{\Delta}, \ \forall \nu = 1, \ldots, m, \tag{3}$$

and $\delta[\mathbf{x}](n) = [\delta[x_1](n), \ldots, \delta[x_m](n)]^T$. A δ -system is stable, if and only if the following condition on the eigenvalues λ_i^{δ} of the matrix A^{δ} is satisfied [1]:

$$| \lambda_i^{\delta} - \Delta^{-1} | < \Delta^{-1}, \quad i = 1, \cdots, m.$$

Therefore a stable system matrix cannot be defective, i.e. it cannot have a zero eigenvalue.

The actual implementation of (1) and (2) in FXP format gives rise to nonlinear quantization operations that occur at various locations depending on the hardware realization.

Eqn. (1) can be implemented either by using single wordlength accumulators (creating a quantization error after each multiplication) or by using double wordlength accumulators (creating a quantization error only after summation). We will only consider the latter option since practically all modern DSP machines offer double precision accumulators.

Eqn. (1) can then be written as

$$\delta[\mathbf{x}](n) = Q\{A^{\delta}\mathbf{x}(n)\}, \tag{4}$$

where Q is a vector-valued quantization nonlinearity of the form

$$Q\{\mathbf{x}\} = \begin{pmatrix} Q\{x_1\} \\ \vdots \\ Q\{x_m\} \end{pmatrix}. \tag{5}$$

Here, $Q\{x_\nu\}$ can denote magnitude truncation, two's complement truncation, or rounding.

Eqn. (2) can be implemented in two different ways:

$$\mathbf{x}(n+1) = \mathbf{x}(n) + Q\{\Delta \cdot \delta[\mathbf{x}](n)\}, \tag{6}$$

or

$$\mathbf{x}(n+1) = Q\{\mathbf{x}(n) + \Delta \cdot \delta[\mathbf{x}](n)\}. \tag{7}$$

Eqn. (6) corresponds to quantization after multiplication while (7) corresponds to quantization after summation. In contrast to (1), for equation (2), it is not clear which of the two quantization schemes in (6) and (7) is preferable. We will therefore consider both possibilities.

Throughout this paper, we will use the following definition of stability:

Definition. The discrete-time system in (4,6) or (4,7) is globally asymptotically stable if and only if, for any initial condition $\mathbf{x}(0)$, the state vector $\mathbf{x}$ asymptotically reaches zero, that is, $\mathbf{x}(n) \to \mathbf{0}$ for $n \to \infty$.

Comment. Since the FXP systems considered are in fact finite state machines, the condition $\mathbf{x}(n) \to \mathbf{0}$ for $n \to \infty$ may be strengthened to $\mathbf{x}(N) = \mathbf{0}$ for some finite N [5].

The following additional symbols will be used:

l: quantization step size

$\underline{0}, \underline{1}$: Vector with all elements being zero or one, respectively.

$Int(x)$: the largest integer function, i.e. the largest integer smaller than or equal to x.

$\mathcal{D}_\delta^{MT}, \mathcal{D}_\delta^{R}, \mathcal{D}_\delta^{TWO}$: Deadbands in terms of the incremental difference vector for magnitude truncation, rounding and two's complement, respectively.

$\mathcal{D}_x^{MT}, \mathcal{D}_x^{R}, \mathcal{D}_x^{TWO}$: Deadbands in terms of the state vector $\mathbf{x}$ for magnitude truncation, rounding and two's complement truncation, respectively.

$\mathcal{A}_\delta^{MT}, \mathcal{A}_\delta^{R}, \mathcal{A}_\delta^{TWO}$: corresponding deadband for the unquantized difference vector.

$\mathcal{H}_L^{MT}$: largest hypercube embedded in $\mathcal{D}_x^{MT}$.

$\mathcal{H}_U^{MT}$: smallest hypercube embedding $\mathcal{D}_x^{MT}$.

III. NECESSARY CONDITIONS FOR GLOBAL ASYMPTOTIC STABILITY

III.1 DC Limit Cycles

First, we will consider the system described by (4,6). From the definition for global asymptotic stability as stated in the previous section, it is necessary that

$$Q\{\Delta \cdot \delta[\mathbf{x}](n)\} \neq \mathbf{0}, \quad \text{for any} \quad \mathbf{x}(n) \neq \mathbf{0}. \tag{8}$$

This is just one of a finite set of conditions that is required to ensure global asymptotic stability of a FXP implementation of a linearly stable system [5].

The following theorem on global asymptotic stability of delta-operator formulated discrete time systems provides conditions on the sampling time:

Theorem 1. A necessary condition for global asymptotic stability of the δ-operator formulated discrete-time system in (4,6) is $\Delta \geq 0.5$ for rounding and $\Delta \geq 1$ for truncation.

Proof: At first, we will address the case of magnitude rounding: The necessary condition

for global asymptotic stability (8) is violated, if

$$| \Delta \cdot \delta[x_\nu](n) | < \frac{l}{2} \quad for \quad \nu = 1, \cdots, m. \tag{9}$$

and $\delta[\mathrm{x}](n) \neq \mathbf{0}$. With

$$\delta[x_\nu](n) = l \quad for \quad \nu = 1, \cdots, m, \tag{10}$$

we can rewrite (9) as

$$\Delta < \frac{1}{2}. \tag{11}$$

If the sampling time is chosen according to (11), then condition (9) is satisfied and hence, the system will exhibit a period one limit cycle. Therefore, in order to avoid a period one limit cycle we require

$$\Delta \geq \frac{1}{2} \tag{12}$$

(Additional constraints will have to be imposed on Δ in order to guarantee the absence of limit cycles with a period other than one.) This proves the Theorem for rounding.

In the case of magnitude truncation, equation(9) becomes:

$$| \Delta \cdot \delta[x_\nu](n) | < l \qquad for \quad \nu = 1, \cdots, m \tag{13}$$

with $\delta[\mathrm{x}](n) \neq \mathbf{0}$ With (13) and (10), one arrives at the following condition, which excludes period one limit cycles:

$$\Delta \geq 1 \tag{14}$$

For two's complement truncation, equation (9) takes the form:

$$0 \leq \Delta \cdot \delta[x_\nu](n) < l \tag{15}$$

Together with (10), the above equation also results in (14), which proves the Theorem.

The above theorem shows that high-speed δ-operator formulated implementations that possess a small sampling time cannot be realized limit cycle free in FXP format! Since the advantages of delta-operator systems with respect to coefficient sensitivity and quantization

noise require a short sampling time much smaller than one, this requirement cannot be met if limit cycles have to be avoided.

A second necessary condition for the system in {(4), (6)} can be obtained by noting that

$$\delta[\mathbf{x}](n) = \mathbf{0} \tag{16}$$

can occur in (4) even though the state vector $\mathbf{x}(n) \neq \mathbf{0}$.

Therefore, for magnitude rounding, no nonzero state vector $\mathbf{x}(n)$ that belongs to the quantization lattice and satisfies

$$-\begin{pmatrix} \frac{\ell}{2} \\ \vdots \\ \frac{\ell}{2} \end{pmatrix} < A^{\delta} \cdot \mathbf{x}(n) < +\begin{pmatrix} \frac{\ell}{2} \\ \vdots \\ \frac{\ell}{2} \end{pmatrix} \tag{17}$$

may be allowed to exist. In (17), the inequality has to hold elementwise.

Equation (17) has the following geometric interpretation:

Each of the resulting m inequalities can be geometrically interpreted in the state space as the intersection of two half spaces in $\Re^m$. These intersections are symmetric about the origin and have parallel boundaries. The normal vector to the boundaries is given by the particular row vector of A^{δ}. Only if the intersection of *all* such m half spaces contains at least one nonzero point in $\Re^m$ on the quantization lattice, will there exist a nonzero state vector that is an equilibrium point of the system due to equation (16). Since we only consider A^{δ} matrices, which are stable, the system matrix A^{δ} is always invertible. One can therefore rewrite (1) to obtain a sufficient condition for the existence of non-zero state vectors, which are equilibrium points due to equation (16):

$$\mathbf{x}(n) = (A^{\delta})^{-1}\delta[\mathbf{x}](n) \quad with \qquad \delta[\mathbf{x}](n) \in (-l/2, l/2)^m \tag{18}$$

In order to obtain bounds for each of the components of $\mathbf{x}(n)$ we use the infinity norm:

$$\| \mathbf{x}(n) \|_{\infty} \leq \| (A^{\delta})^{-1} \|_{\infty} \| \delta[\mathbf{x}](n) \|_{\infty} < \| (A^{\delta})^{-1} \|_{\infty} \frac{l}{2} \tag{19}$$

The perallelepiped described by (18) is therefore imbedded in the hypercuboid described by (19). If (19) does not permit any points $\mathbf{x(n)}$ of the sampling lattice, instability due to (16) cannot occur. From (19), this is the case if

$$\| (A^{\delta})^{-1} \|_{\infty} < 2. \tag{20}$$

Eqn. (16) can also be interpreted from an eigenvalue/eigenvector viewpoint. In high-speed digital filters where the sampling frequency is typically much higher than the bandwidth of the processed signal, the eigenvalues of a q-operator implementation cluster around the point $z = 1$ [1]. The corresponding δ-operator implementation for large sampling times has eigenvalues clustered around zero. However, as the sampling time becomes small, these eigenvalues move towards the eigenvalues of the underlying continuous-time system [1]. In other words, for large sampling times, the system matrix will be ill-conditioned, that is, vectors $\mathbf{x}(n) \neq \mathbf{0}$ exist such that $A^{\delta} \cdot \mathbf{x}(n)$ is close to the zero vector. According to (16), this is likely to cause a DC limit cycle. For small sampling times, this problem may not occur; however, in this case, the conditions in Theorem 1 are not satisfied and the system is already known to produce limit cycles.

In the case of the remaining two quantization schemes, the inequalities corresponding to (17) are given below: For two's complement truncation,

$$\mathbf{0} \leq A^{\delta} \cdot \mathbf{x}(n) < \begin{pmatrix} \ell \\ \vdots \\ \ell \end{pmatrix}, \ \mathbf{x}(n) \neq \mathbf{0}, \tag{21}$$

and, for magnitude truncation,

$$-\begin{pmatrix} \ell \\ \vdots \\ \ell \end{pmatrix} < A^{\delta} \cdot \mathbf{x}(n) < +\begin{pmatrix} \ell \\ \vdots \\ \ell \end{pmatrix}, \ \mathbf{x}(n) \neq \mathbf{0}. \tag{22}$$

Again, the above inequalities have to interpreted elementwise. The embedding hypercubes can be constructed for the perallelepiped in (21) and (22) in a similar fashion as for rounding in (18).

So far, we only addressed the system described by (4,6). A similar analysis can be conducted for the system in (4,7). Since (4) is common to both realizations, equations

(17,21,22) are still valid and provide conditions under which the finite difference is quantized to zero and a DC limit cycle is produced. We will now briefly discuss necessary conditions for global asymptotic stability obtained from (7).

A period one limit cycle exists, if the condition

$$\mathbf{x} = Q(\mathbf{x} + \Delta\delta[\mathbf{x}](n)) \tag{23}$$

is satisfied for $\mathbf{x} \neq \mathbf{0}$. Using a similar argument as in the proof of Theorem 1, for rounding, equation (23) is satisfied if:

$$-\frac{l}{2} \leq \Delta\delta[x_\nu](n) < \frac{l}{2} \qquad for \quad x_\nu > 0 \tag{24}$$

$$-\frac{l}{2} < \Delta\delta[x_\nu](n) \leq \frac{l}{2} \qquad for \quad x_\nu < 0 \tag{25}$$

$$-\frac{l}{2} < \Delta\delta[x_\nu](n) < \frac{l}{2} \qquad for \quad x_\nu = 0 \tag{26}$$

$$\nu = 1, \cdots, m$$

Therefore

$$\Delta > \frac{1}{2} \tag{27}$$

is required to exclude period one limit cycles.

For magnitude truncation, (23) is satisfied, if

$$0 \leq \Delta\delta[x_\nu](n) < l \qquad for \quad x_\nu > 0 \tag{28}$$

$$-l < \Delta\delta[x_\nu](n) \leq 0 \qquad for \quad x_\nu < 0 \tag{29}$$

$$-l < \Delta\delta[x_\nu](n) < l \qquad for \quad x_\nu = 0 \tag{30}$$

$$\nu = 1, \cdots, m$$

In the case of two's complement truncation, the condition for a DC limit cycle is simply given by

$$0 \leq \Delta\delta[x_\nu](n) < l, \qquad \nu = 1, \cdots, m. \tag{31}$$

The conditions (28-30) and (31) again result in the condition $\Delta \geq 1$ for the absence of period one limit cycles.

We therefore obtain almost the same conclusion as for the previously considered system:

$$\Delta > \frac{1}{2} \quad \text{for magnitude rounding;}$$

$$\Delta \geq 1 \quad \text{for truncation.}$$

Therefore, Theorem 1 also holds for the system representation in $\{(4),(7)\}$, if the condition for rounding is slightly changed to $\Delta > \frac{1}{2}$.

Upto now, we provided necessary conditions for stability of delta-operator fomulated discrete time systems in fixed point arithmetic. Since it has been established, that for small sampling periods, the delta-operator systems always exhibits period one limit cycles, one needs to examine the amplitude of these limit cycles for a given sampling time in order to obtain further insight into the practical impact of this problem. In what follows, bounds on the deadbands will be derived as a function of the A^δ-matrix and the sampling time Δ.

III.2 Deadband Bounds

This subsection provides an answer to the question of the size of the limit cycle amplitudes. Given a sampling time Δ and a system matrix A^δ, bounds for the deadbands as well as the deadband geometry will be described. This will be done in detail for the case of magnitude truncation. For magnitude rounding and two's complement truncation, the results will be stated briefly without proof. Since the results for the system (4,7) are very similar to the results for the system(4,6), this subsection focuses only on the latter.

For each quantization scheme, we will provide the geometry of the deadband in terms of the incremental difference vector as well as the state vector. Two hypercubes, which bound the deadband region from the inside and the outside are also derived for each case.

Theorem 2:

For the system (4,6) implemented in magnitude truncation, the deadband (in terms of

period one limit cycles) in the incremental difference vector space is given by:

$$\mathcal{D}_\delta^{MT} = \{\delta[\mathrm{x}] \mid \quad \| \delta[\mathrm{x}] \|_\infty \leq [Int(\Delta^{-1}) - 1] \cdot l\} \qquad for \quad Int(\Delta^{-1}) = \Delta^{-1} \tag{32}$$

and

$$\mathcal{D}_\delta^{MT} = \{\delta[\mathrm{x}] \mid \quad \| \delta[\mathrm{x}] \|_\infty \leq Int(\Delta^{-1}) \cdot l\} \qquad for \quad Int(\Delta^{-1}) \neq \Delta^{-1}. \tag{33}$$

The corresponding period one limit cycle deadband in the state space is given by

$$\mathcal{D}_x^{MT} = \{\mathrm{x} \mid \mathrm{x} = (A^\delta)^{-1}\delta[\mathrm{x}], \quad \delta[\mathrm{x}] \in \mathcal{A}_\delta^{MT}\} \tag{34}$$

where

$$\mathcal{A}_\delta^{MT} = \{\delta[\mathrm{x}] \mid \quad \| \delta[\mathrm{x}] \|_\infty < [Int(\Delta^{-1}) + 1] \cdot l\} \qquad for \quad Int(\Delta^{-1}) \neq \Delta^{-1} \tag{35}$$

and

$$\mathcal{A}_\delta^{MT} = \{\delta[\mathrm{x}] \mid \quad \| \delta[\mathrm{x}] \|_\infty < Int(\Delta^{-1}) \cdot l\} \qquad for \quad Int(\Delta^{-1}) = \Delta^{-1} \tag{36}$$

Proof:

The proof will be carried out for $Int(\Delta^{-1}) \neq \Delta^{-1}$, since the case $Int(\Delta^{-1}) = \Delta^{-1}$ follows in a similar fashion. From (13), the expression for period one limit cycles can be expressed as

$$\| \Delta\delta[\mathrm{x}](n) \|_\infty < l. \tag{37}$$

Solving (37) for $\delta[\mathrm{x}]$ and considering, that $\delta[\mathrm{x}]$ produced by equation (4) is an integer multiple of the quantization step l, one obtains

$$\| \delta[\mathrm{x}](n) \|_\infty \leq Int(\Delta^{-1}) \cdot l \tag{38}$$

for $Int(\Delta^{-1}) \neq \Delta^{-1}$ which is the hypercube in (33). Now consider the following slightly larger hypercube $\mathcal{A}_\delta^{MT}$ in $\delta[\mathrm{x}]$:

$$\mathcal{A}_\delta^{MT} = \{\delta[\mathrm{x}] \mid \quad \| \delta[\mathrm{x}] \|_\infty < [Int(\Delta^{-1}) + 1]l\} \tag{39}$$

$\mathcal{A}_\delta^{MT}$ describes the open set of all incremental difference vectors, which, after quantization will be mapped into the hypercube $\mathcal{D}_\delta^{MT}$, i.e.

$$Q(\delta[\mathrm{x}]) \in \mathcal{D}_\delta^{MT}, \quad \forall \delta[\mathrm{x}] \in \mathcal{A}_\delta^{MT}.$$

Therefore the deadband in terms of $\mathbf{x}$ can simply be found by determining the set of all $\mathbf{x}$, which satisfy

$$A^{\delta}\mathbf{x} \in \mathcal{A}_{\delta}^{MT}.$$

Since A^{δ} was assumed to be linearly stable, it is also invertible . Therefore the deadband in the state space is obtained by

$$\mathcal{D}_{x}^{MT} = \{\mathbf{x} \mid \mathbf{x} = (A^{\delta})^{-1}\delta[\mathbf{x}], \quad \delta[\mathbf{x}] \in \mathcal{A}_{\delta}^{MT}\}$$

This completes the proof for $Int(\Delta^{-1}) \neq \Delta^{-1}$.

The following Corollary provides the largest hypercube in the state space, which is contained in the perallelepiped $\mathcal{D}_{x}^{MT}$. This result allows to obtain the largest magnitude of state vector components, which can still belong to the deadband. It also provides a simple upper bound on the volume of the deadband.

Corollary 3:

The largest hypercube $\mathcal{H}_{L}^{MT}$ embedded in $\mathcal{D}_{x}^{MT}$ is given by:

$$\mathcal{H}_{L}^{MT} = \{\mathbf{x} \mid \quad \| \mathbf{x}(n) \|_{\infty} < \frac{[Int(\Delta^{-1}) + 1]l}{\| A^{\delta} \|_{\infty}}\} \qquad for \quad Int(\Delta^{-1}) \neq \Delta^{-1} \tag{40}$$

and by

$$\mathcal{H}_{L}^{MT} = \{\mathbf{x} \mid \quad \| \mathbf{x}(n) \|_{\infty} < \frac{Int(\Delta^{-1})l}{\| A^{\delta} \|_{\infty}}\} \qquad for \quad Int(\Delta^{-1}) = \Delta^{-1} \tag{41}$$

Proof:

Assume $Int(\Delta^{-1}) \neq \Delta^{-1}$. From (1) we obtain for the unquantized incremental difference vector:

$$\| \delta[\mathbf{x}] \|_{\infty} \leq \| A^{\delta} \|_{\infty} \| \mathbf{x} \|_{\infty} \tag{42}$$

Since $\mathcal{A}_{\delta}^{MT}$ describes the set of unquantized difference vectors, which after quantization maps into the deadband region $\mathcal{D}_{\delta}^{MT}$, one can use the right side of (42) to ensure, that equation (39) is satisfied and obtain:

$$\| A^{\delta} \|_{\infty} \| \mathbf{x} \|_{\infty} < [Int(\Delta^{-1}) + 1] \cdot l \tag{43}$$

Solving (43) for $\| \mathbf{x} \|_\infty$ produces the desired result. Since $\mathcal{H}_L^{MT}$ is a hypercube centered at the origin, there exists a $\mathbf{x} \in \mathcal{H}_L^{MT}$, such that

$$\| \delta[\mathbf{x}] \|_\infty = \| A^\delta \|_\infty \cdot \| \mathbf{x} \|_\infty . \tag{44}$$

Hence this is the largest such hypercube. The proof for the case $Int(\Delta^{-1}) = \Delta^{-1}$ follows from (43) in a similar fashion.

The next Corollary provides the smallest hypercube in the state space, which still contains $\mathcal{D}_x^{MT}$. This provides a lower bound on the volume of the deadband:

Corollary 4:

The smallest hypercube $\mathcal{H}_U^{MT}$ containing $\mathcal{D}_x^{MT}$ is given by

$$\mathcal{H}_U^{MT} = \{\mathbf{x} \mid \quad \| \mathbf{x} \|_\infty < \| (A^\delta)^{-1} \|_\infty \, (Int(\Delta^{-1}) + 1) \cdot l\} \qquad for \quad Int(\Delta^{-1}) \neq \Delta^{-1} \tag{45}$$

and

$$\mathcal{H}_U^{MT} = \{\mathbf{x} \mid \quad \| \mathbf{x} \|_\infty < \| (A^\delta)^{-1} \|_\infty \, Int(\Delta^{-1}) \cdot l\} \qquad for \quad Int(\Delta^{-1}) = \Delta^{-1} \tag{46}$$

Proof:

At first consider the case $Int(\Delta^{-1}) \neq \Delta^{-1}$: From (1) we have for the unquantized state vector:

$$\mathbf{x} = (A^\delta)^{-1} \delta[\mathbf{x}] \tag{47}$$

Taking norms and using the inequality in (35), we obtain the following open hypercube, which contains $\mathcal{D}_x^{MT}$:

$$\| \mathbf{x} \|_\infty \leq \| (A^\delta)^{-1} \|_\infty \| \delta[\mathbf{x}] \|_\infty < \| (A^\delta)^{-1} \|_\infty \, [Int(\Delta^{-1}) + 1] \cdot l$$

Since $\mathcal{D}_\delta^{MT}$ is a hypercube centered at the origin, there exists a $\delta[\mathbf{x}]$, such that

$$\| (A^\delta)^{-1} \|_\infty \| \delta[\mathbf{x}] \|_\infty = \| \mathbf{x} \|_\infty .$$

Hence $\mathcal{H}_U^{MT}$ is the smallest such hypercube. The proof for the case $Int(\Delta^{-1}) = \Delta^{-1}$ is identical and requires the use of (36) instead of (35).

Remarks:

1. Since $\| (A^\delta)^{-1} \| \cdot \| (A^\delta) \| \geq 1$, we have $\mathcal{H}_L^{MT} \subset \mathcal{H}_U^{MT}$. For matrices which satisfy

$$\| (A^\delta)^{-1} \|_\infty \cdot \| A^\delta \|_\infty = 1 \tag{48}$$

the two hypercubes are identical and coincide with the deadband region $\mathcal{D}_x^{MT}$, i.e. $\mathcal{H}_U^{MT} = \mathcal{H}_L^{MT} = \mathcal{D}_x^{MT}$.

2. $\mathcal{D}_\delta^{MT}$ is a closed m-D hypercube, centered around the origin. Its boundary coincides with points of the quantization lattice. The faces of the hypercube are orthogonal to the corresponding axis of the incremental difference vector space.

3. $\mathcal{D}_x^{MT}$ is an open parallelepiped. $\mathcal{D}_x^{MT}$ describes the deadband for period one limit cycles in terms of the state vector.

4. The total deadband includes the region $\mathcal{D}_\delta^{MT}$ ($\mathcal{D}_x^{MT}$) for the incremental difference vector (the state vector.)

5. A useful measure of the deadband size in terms of the state vector $\mathbf{x}$ is the volume in the state space. The 'volume' Vol_δ of the deadband in $\delta[\mathrm{x}]$ is easily computable due to the hypercube geometry. From (47) we obtain for the volume Vol_x in the state space of $\mathbf{x}$:

$$Vol_x = det((A^\delta)^{-1}) \cdot Vol_\delta \tag{49}$$

6. Given a realization, increasing the sampling rate (Δ^{-1}) will result in a larger deadband.

The relationships for the deadband of quantization schemes other than magnitude truncation are given below:

Magnitude Rounding:

$$\mathcal{D}_\delta^R = \{\delta[\mathrm{x}] \mid \quad \| \delta[\mathrm{x}] \|_\infty \leq [Int(\frac{1}{2}\Delta^{-1}) - 1] \cdot l\} \qquad for \quad Int(\frac{1}{2}\Delta^{-1}) = \frac{1}{2}\Delta^{-1} \tag{50}$$

and

$$\mathcal{D}_\delta^R = \{\delta[\mathbf{x}] \mid \quad \| \delta[\mathbf{x}] \|_\infty \leq Int(\frac{1}{2}\Delta^{-1}) \cdot l\} \qquad for \quad Int(\frac{1}{2}\Delta^{-1}) \neq \frac{1}{2}\Delta^{-1} \tag{51}$$

For the deadband in terms of the state vector we have:

$$\mathcal{D}_x^R = \{\mathbf{x} \mid \mathbf{x} = (A^\delta)^{-1}\delta[\mathbf{x}], \delta[\mathbf{x}] \in \mathcal{A}_\delta^R\} \tag{52}$$

where

$$\mathcal{A}_\delta^R = \{\delta[\mathbf{x}] \mid \quad \| \delta[\mathbf{x}] \|_\infty < [Int(\frac{1}{2}\Delta^{-1}) - \frac{1}{2}] \cdot l\} \qquad for \quad Int(\frac{1}{2}\Delta^{-1}) = \frac{1}{2}\Delta^{-1} \tag{53}$$

and

$$\mathcal{A}_\delta^R = \{\delta[\mathbf{x}] \mid \quad \| \delta[\mathbf{x}] \|_\infty < [Int(\frac{1}{2}\Delta^{-1}) + \frac{1}{2}] \cdot l\} \qquad for \quad Int(\frac{1}{2}\Delta^{-1}) \neq \frac{1}{2}\Delta^{-1} \tag{54}$$

Two's Complement Truncation:

$$\mathcal{D}_\delta^{TWO} = \{\delta[\mathbf{x}] \mid \underline{0} \leq \delta[\mathbf{x}] \leq \underline{1} \cdot [Int(\Delta^{-1}) - 1] \cdot l\} \qquad for \quad Int(\Delta^{-1}) = \Delta^{-1} \tag{55}$$

and

$$\mathcal{D}_\delta^{TWO} = \{\delta[\mathbf{x}] \mid \underline{0} \leq \delta[\mathbf{x}] \leq \underline{1} \cdot Int(\Delta^{-1}) \cdot l\} \qquad for \quad Int(\Delta^{-1}) \neq \Delta^{-1}. \tag{56}$$

For the deadband in terms of the state vector we have:

$$\mathcal{D}_x^{TWO} = \{\mathbf{x} \mid \mathbf{x} = (A^\delta)^{-1}\delta[\mathbf{x}], \quad \delta[\mathbf{x}] \in \mathcal{A}_\delta^{TWO}\} \tag{57}$$

where

$$\mathcal{A}_\delta^{TWO} = \{\delta[\mathbf{x}] \mid \underline{0} \leq \delta[\mathbf{x}] < \underline{1} \cdot Int(\Delta^{-1}) \cdot l\} \qquad for \quad Int(\Delta^{-1}) = \Delta^{-1} \tag{58}$$

and

$$\mathcal{A}_\delta^{TWO} = \{\delta[\mathbf{x}] \mid \underline{0} \leq \delta[\mathbf{x}] < \underline{1} \cdot [Int(\Delta^{-1}) + 1] \cdot l\} \qquad for \quad Int(\Delta^{-1}) \neq \Delta^{-1}. \tag{59}$$

In the above set definitions, all inequalities are to be interpreted elementwise, i.e. $\mathbf{x} \leq \mathbf{y}$ with $\mathbf{x}, \mathbf{y} \in \mathcal{R}^m$ means $x_i \leq y_i, \quad i = 1, \cdots, m$. Furthermore, the notation $\underline{0}, \underline{1}$ stands for the zero vector and the vector with component values of one, respectively.

IV. THE M-D CASE

IV.1 Additional Notation for the m-D Case

The m-D Roesser model has the following δ-operator formulation [7]:

$$\begin{bmatrix} \delta^{(1)}[\mathbf{x}^{(1)}](\mathbf{n}) \\ \vdots \\ \delta^{(m)}[\mathbf{x}^{(m)}](\mathbf{n}) \end{bmatrix} = \begin{bmatrix} A_{11}^{\delta} & \cdots & A_{1m}^{\delta} \\ \vdots & \ddots & \vdots \\ A_{m1}^{\delta} & \cdots & A_{mm}^{\delta} \end{bmatrix} \begin{bmatrix} \mathbf{x}^{(1)}(\mathbf{n}) \\ \vdots \\ \mathbf{x}^{(m)}(\mathbf{n}) \end{bmatrix} + \begin{bmatrix} B_1^{\delta} \\ \vdots \\ B_m^{\delta} \end{bmatrix} \mathbf{u}(\mathbf{n}); \tag{60}$$

$$\begin{bmatrix} q^{(1)}[\mathbf{x}^{(1)}](\mathbf{n}) \\ \vdots \\ q^{(m)}[\mathbf{x}^{(m)}](\mathbf{n}) \end{bmatrix} = \begin{bmatrix} \mathbf{x}^{(1)}(\mathbf{n}) \\ \vdots \\ \mathbf{x}^{(m)}(\mathbf{n}) \end{bmatrix} + \Delta \cdot \begin{bmatrix} \delta^{(1)}[\mathbf{x}^{(1)}](\mathbf{n}) \\ \vdots \\ \delta^{(m)}[\mathbf{x}^{(m)}](\mathbf{n}) \end{bmatrix}. \tag{61}$$

The input-state equations in (60) and (61) describe a first hyper-quadrant causal m-D system with a uniform sampling period of Δ in all directions. The operators $q^{(i)}$ and $\delta^{(i)}$ represent the shift- and delta-operator in the direction specified by the axis n_i. In particular

$$q^{(i)}[\mathbf{x}^{(i)}](\mathbf{n}) = \mathbf{x}^{(i)}(n_1, \ldots, n_{i-1}, n_i + 1, n_{i+1}, \ldots, n_m) \tag{62}$$

$$\delta^{(i)}[\mathbf{x}^{(i)}](\mathbf{n}) = \frac{1}{\Delta}(\mathbf{x}^{(i)}(n_1, \ldots, n_{i-1}, n_i + 1, n_{i+1}, \ldots, n_m) - \mathbf{x}^{(i)}(\mathbf{n})). \tag{63}$$

Here, $(\mathbf{n}) \doteq (n_1, \ldots, n_m)$ denotes a point in the first hyper-quandrant, $\mathbf{x}^{(i)}(\mathbf{n})$ is the portion of the state vector propagating in the direction specified by the axis n_i, $u(\mathbf{n})$ is the m-D input vector, and A_{ij}^{δ} and B_i^{δ}, for $i = 1, \ldots, m$, $j = 1, \ldots, m$, are the submatrices of the system and input matrices, respectively.

If (60) is realized in fixed-point arithmetic, it takes the following form under zero-input conditions:

$$\begin{bmatrix} \delta^{(1)}[\mathbf{x}^{(1)}](\mathbf{n}) \\ \vdots \\ \delta^{(m)}[\mathbf{x}^{(m)}](\mathbf{n}) \end{bmatrix} = \mathbf{Q}\left\{ \begin{bmatrix} A_{11}^{\delta} & \cdots & A_{1m}^{\delta} \\ \vdots & \ddots & \vdots \\ A_{m1}^{\delta} & \cdots & A_{mm}^{\delta} \end{bmatrix} \begin{bmatrix} \mathbf{x}^{(1)}(\mathbf{n}) \\ \vdots \\ \mathbf{x}^{(m)}(\mathbf{n}) \end{bmatrix} \right\} \tag{64}$$

Equation (64) assumes quantization after summation; since practically all modern DSP machines implement this quantization scheme, we only consider this format. The

vector-valued quantization nonlinearity $\mathbf{Q}\{\cdot\}$ may represent any one of the conventional schemes, viz., magnitude truncation, magnitude rounding, two's complement truncation, and two's complement rounding.

Equation (61) can be implemented in two different forms:

$$\begin{bmatrix} q^{(1)}[\mathbf{x}^{(i)}](\mathbf{n}) \\ \vdots \\ q^{(m)}[\mathbf{x}^{(m)}](\mathbf{n}) \end{bmatrix} = \begin{bmatrix} \mathbf{x}^{(1)}(\mathbf{n}) \\ \vdots \\ \mathbf{x}^{(m)}(\mathbf{n}) \end{bmatrix} + \mathbf{Q}\left\{ \Delta \cdot \begin{bmatrix} \delta^{(1)}[\mathbf{x}^{(1)}](\mathbf{n}) \\ \vdots \\ \delta^{(m)}[\mathbf{x}^{(m)}](\mathbf{n}) \end{bmatrix} \right\} \tag{65}$$

or

$$\begin{bmatrix} q^{(1)}[\mathbf{x}^{(1)}](\mathbf{n}) \\ \vdots \\ q^{(m)}[\mathbf{x}^{(m)}](\mathbf{n}) \end{bmatrix} = \mathbf{Q}\left\{ \begin{bmatrix} \mathbf{x}^{(1)}(\mathbf{n}) \\ \vdots \\ \mathbf{x}^{(m)}(\mathbf{n}) \end{bmatrix} + \Delta \cdot \begin{bmatrix} \delta^{(1)}[\mathbf{x}^{(1)}](\mathbf{n}) \\ \vdots \\ \delta^{(m)}[\mathbf{x}^{(m)}](\mathbf{n}) \end{bmatrix} \right\}. \tag{66}$$

Equation (65) corresponds to quantization after multiplication, whereas (66) corresponds to quantization after addition. In contrast to (60), for (61), it is not obvious which of the two forms stated above is preferable.

The following definition for asymptotic stability [8] will be used throughout this paper.

Definition. An m-D first hyper-quadrant causal discrete-time system is asymptotically stable under all finitely extended bounded input signals $u(\mathbf{n})$ where

$$|u(\mathbf{n})| \leq S, \quad \text{for} \quad n_1 + \cdots + n_m \leq D; \tag{67}$$

$$u(\mathbf{n}) = 0, \quad \text{for} \quad n_1 + \cdots + n_m > D, \tag{68}$$

if all the states of the m-D discrete-time system asymptotically reach zero for $n_1 + \cdots + n_m \rightarrow \infty$. Here, $n_\nu \geq 0$, $\nu = 1, \ldots, m$, S is a nonnegative real number, and D is a positive integer.

Since the fixed-point systems considered are in fact finite state machines, the condition

$$\begin{pmatrix} \mathbf{x}^{(1)}(\mathbf{n}) \\ \vdots \\ \mathbf{x}^{(m)}(\mathbf{n}) \end{pmatrix} \rightarrow \mathbf{0},$$

for $n_1 + \cdots + n_m \to \infty$, $n_\nu \geq 0$, $\nu = 1, \ldots, m$, can be strengthened to

$$\begin{pmatrix} \mathbf{x}^{(1)}(\mathbf{n}) \\ \vdots \\ \mathbf{x}^{(m)}(\mathbf{n}) \end{pmatrix} = \mathbf{0},$$

for all points $n_1 + \cdots + n_m \geq c$, $n_\nu \geq 0$, $\nu = 1, \ldots, m$, where c is some finite integer.

IV.2 Necessary Conditions for Global Asymptotic Stability

In this section, we present some necessary conditions for stability of a first hyper-quadrant causal m-D discrete-time system represented in its Roesser local state-space model in (60,61). These necessary conditions are formulated in terms of 1-D conditions. This theorem follows directly from a result in [6] which was formulated for q-operator implemented discrete-time systems. The proof of the theorem rests on the fact that a first hyper-quadrant m-D system can be described by a 1-D system for those locations that are along the m coordinate axes of the boundary of the hyper-quadrant. Reformulating the result in [6] for δ-operator systems produces the following theorem:

Theorem 5.

(a) A necessary condition for global asymptotic stability of the system in (64,65) is that each of the following 1-D systems in (69,70) is globally asymptotically stable:

$$\delta^{(i)}[\mathbf{x}^{(i)}](n_i) = \mathbf{Q}\left\{[A_{ii}^{\delta}]\mathbf{x}^{(i)}(n_i)\right\}; \tag{69}$$

$$q^{(i)}[\mathbf{x}^{(i)}](n_i) = \mathbf{x}^{(i)}(n_i) + \mathbf{Q}\left\{\Delta \cdot \delta^{(i)}[\mathbf{x}^{(i)}](n_i)\right\}, \tag{70}$$

where $i = 1, \ldots, m$.

(b) A necessary condition for global asymptotic stability of the system in (64,66) is that each of the following in 1-D systems in (71,72) is globally asymptotically stable:

$$\delta^{(i)}[\mathbf{x}^{(i)}](n_i) = \mathbf{Q}\left\{[A_{ii}^{\delta}]\mathbf{x}^{(i)}(n_i)\right\}; \tag{71}$$

$$q^{(i)}[\mathbf{x}^{(i)}](n_i) = \mathbf{Q}\left\{\mathbf{x}^{(i)}(n_i) + \Delta \cdot \delta^{(i)}[\mathbf{x}^{(i)}](n_i)\right\}, \tag{72}$$

where $i = 1, \ldots, m$.

Proof. For a detailed proof, and generalizations to higher sub-dimensional systems, the reader is referred to [6].

Theorem 5 can be viewed as an extension of the concept of practical BIBO stability to asymptotic stability of nonlinear systems. It is particularly useful in proving instability in m-D nonlinear systems.

We can now combine Theorem 1 and Theorem 5 to formulate a necessary condition for stability of m-D first hyper-quadrant causal δ-operator formulations of the generalized Roesser model.

Corollary 6.

(a) A necessary condition for global asymptotic stability of the m-D systems in (64,65) is

$$\Delta \geq 0.5, \quad \text{for magnitude rounding;}$$

$$\Delta \geq 1, \quad \text{for truncation.}$$

(b) A necessary condition for global asymptotic stability of the $m-$D system in (64,66) is

$$\Delta > 0.5, \quad \text{for magnitude rounding;}$$

$$\Delta \geq 1, \quad \text{for truncation.}$$

Proof. The proof follows from Theorems 1 and 5.

Remarks:

1. Corollary 6 is also essentially applicable to the case where the sampling time varies with the direction of propagation. In the case of the system description (64,65), the inequalities in Corollary 6 would have to be replaced by

$$\Delta_i \geq 0.5, \quad \text{for magnitude rounding;}$$

$$\Delta_i \geq 1, \quad \text{for truncating,}$$

for $i = 1, \ldots, m$. The conditions for the system (64,66) are analogous.

2. Our analysis is limited to the zero-input case for which DC limit cycles along the axis were used to derive conditions for non-convergence. If one includes other types of limit cycles in the analysis or even response types, which are not periodic and are known to exist only in the m-D case, the requirements for Δ may become even more severe.

3. Corollary 6 shows that fixed-point implementations of 1-D and m-D δ-operator systems *cannot be realized limit cycle free, if good coefficient sensitivity and quantization noise measures have to be achieved.*

V. CONCLUSION

In this paper, it was shown that fixed-point implementations of 1-D and m-D δ-operator systems are not limit cycle free even if the underlying linear system is stable and the sampling time is chosen small. This non-convergent behavior can be explained by the quantization of the δ-term to zero which leaves the state vector unchanged. The smaller the sampling time, the more severe this effect. The size of the deadband increases with a decreasing sampling time. Therefore, the practical value of δ-operators for fixed-point implementations of 1-D and m-D systems is questionable. There are however indications that this effect is much less severe in floating-point implementations.

δ-operator implemented discrete-time systems represent a class of systems where the quantization noise at the output can be small compared to other realizations. However, as was shown above, such realizations will invariably exhibit limit cycles, which are highly correlated quantization noise. Therefore, in this case, typical measures for quantization noise are of very limited use for obtaining any insight into the likelihood of limit cycles and vice versa.

ACKNOWLEDGEMENT

This work was supported by two grants from the Office of Naval Research (ONR): N 00014-94-1-0454 and N 00014-94-1-0387.

REFERENCES

[1] G.C. Goodwin, R.H. Middleton, and H.V. Poor, "High-speed digital signal processing and control," *Proceedings of the IEEE,* vol. 80, no. 2, pp. 240-259, Feb. 1992.

[2] R.H. Middleton and G.C. Goodwin, "Improved finite wordlength characteristics in digital control using delta operators," *IEEE Transactions on Automatic Control,* vol. 31, pp. 1015-1021, Nov. 1986.

[3] G.Li and M. Gevers, "Comparative study of finite wordlength effects in shift and delta operator parameterization," *Proceedings of the IEEE Conference on Decision and Control (CDC'90),* vol. 2, pp. 954-959, Honolulu, HI, 1990.

[4] G. Li and M. Gevers, "Roundoff noise minimization using delta-operator realizations", *IEEE Transactions on Signal Processing*, Vol. 41, No. 2, pp. 629-637, Feb. 1993.

[5] P. H. Bauer and L. J. Leclerc, "A computer-aided test for the absence of limit cycles in fixed point digital filters", *IEEE Transactions on Signal Processing*, Vol. 39, No. 11, pp. 2400-2410, Nov. 1991.

[6] P. Bauer, "Low-dimensional conditions for global asymptotic stability of m-D nonlinear digital filters," *1994 IEEE International Symposium on Circuits and Systems,* London, England, pp. 2.461-2.464.

[7] K. Premaratne, J. Suarez, M. Ekanayake, P.H. Bauer, "Two-dimensional delta-operator formulated discrete time systems: State space realization and its coefficient sensitivity properties", *Proceedings of the 37th Midwest Symposium on Circuits and Systems*, Aug. 1994, Lafayette, LA

[8] P. Bauer, "Finite wordlength effects in $m-$D digital filters with singularities on the stability boundary," *IEEE Transactions on Signal Processing* vol. 40, no. 4, pp. 894-900, April 1994.

IEEE Transactions on Signal Processing, in review, January 1995. apr0194*.tex

Two-Dimensional Delta-Operator Formulated Discrete-Time Systems: State-Space Realization and Its Coefficient Sensitivity Properties

K. Premaratne, *Senior Member, IEEE,* M.M. Ekanayake, *Student Member, IEEE,* J. Suarez, *Student Member, IEEE,* and P.H. Bauer, *Member, IEEE*

Abstract. **Recently, delta-operator based implementation of one-dimensional discrete-time systems has been the focus of considerable research activity. This is due mainly to its superior finite wordlength properties and the possibility of providing a unified treatment of both continuous- and discrete-time systems. In this paper, we investigate delta-operator formulated implementation of two-dimensional discrete-time systems. For this purpose, the δ-operator based counterpart to the Roesser local state-space realization is introduced. Reachability and observability gramians and the notion of a balanced realization for such a model are defined, and their computation is addressed. Coefficient sensitivity properties of the resulting implementations (under both fixed- and floating-point arithmetic), and conditions under which the proposed delta-operator based model is superior, are also derived.**

EDICS Number: SP 4.1

Corresponding Author:
KAMAL PREMARATNE
Department of Electrical and Computer Engineering
University of Miami
1251 Memorial Drive #EB406
Coral Gables, FL 33146 USA.

Tel: +1(305)284 4051
Fax: +1(305)284 4044
email: kprema@umiami.ir.miami.edu

Manuscript received—

K. Premaratne, M.M. Ekanayake, and J. Suarez are with the Department of Electrical and Computer Engineering, University of Miami, 1251 Memorial Drive #EB406, Coral Gables, FL 33146 USA. P.H. Bauer is with the Department of Electrical Engineering, University of Notre Dame, Notre Dame, IN 46556 USA.

K.P. and P.H.B. gratefully acknowledge the support provided by the U.S. Office of Naval Research (ONR) through grants N00014-94-1-0454 and N00014-94-1-0387, respectively.

I. Introduction

Current interest in δ-systems is due mainly to two reasons: (a) δ-systems provide superior roundoff noise [1-2] and coefficient sensitivity [3-4] properties, and (b) δ-operator makes it possible to treat both continuous-time (CT) and discrete-time (DT) systems in a unified manner since it yields the differential operator as a limiting case [5-6].

Hence, implementation of two-dimensional (2-D) and multi-dimensional (m-D) systems using the δ-operator can be expected to provide digital filters that perform better in a shorter wordlength environment. If this is the case, such implementations can find widespread use in high performance, real-time applications, where fast sampling and/or shorter wordlength are desired. In such cases, traditional q-operator implementations perform poorly [7].

With this in mind, research directed towards developing models for 2-D and m-D δ-systems is warranted. This paper presents a local state-space (s.s.) model that is the counterpart to the well known q-operator based Roesser model [8]. We also define the notions of gramians and balanced (BL) realization, and address their computation. With these tools in hand, we then investigate coefficient sensitivity properties of this model. Indeed, implementation of 2-D and m-D systems using this *Roesser δ-model,* under mild conditions, is shown to provide superior coefficient sensitivity compared to the more conventional implementation of *Roesser q-model.* As usual, for notational simplicity, we concentrate only on the 2-D case, the extension to the m-D case being quite straight-forward.

The paper is organized as follows: Section II provide the nomenclature, some preliminary material, and a brief review of relevant results. Section III contains the development of the Roesser δ-model and some important system theoretic notions. In particular, after establishing the connection between the gramians of one-dimensional (1-D) q- and δ-systems, we define the notion of gramians for 2-D δ-systems. Relationship between these and those corresponding to 2-D q-systems, gramians, notion of a BL realization, and its computation are then presented. Investigation of coefficient sensitivity of the δ-model is in Section IV. Addressing the more general multi-input multi-output (MIMO) case, for this purpose, two sensitivity measures—applicable for fixed-point (FXP) and floating-point (FLP) arithmetic schemes—are proposed. Conditions under which the proposed δ-model offers superior coefficient sensitivity are also derived. Section V contains an example. Section VI is reserved for concluding remarks.

II. Nomenclature and Preliminaries

2.2. Nomenclature

$\Re$, $\Im$, $\aleph$	Reals, complex numbers, nonnegative integers.
$\Re^{q\times p}$, $\Im^{q\times p}$	Set of matrices of size $q \times p$ over $\Re$ and $\Im$.
$\Re[w]_n$, $\Im[w]_n$	Set of univariate polynomials of degree n (with respect to indeterminate $w \in \Im$) over $\Re$ and $\Im$.
$\Re(w)_n$	Set of rational univariate polynomials of degree n (with respect to indeterminate $w \in \Im$) over $\Re$.
$\Re[w_h]_{n_h}[w_v]_{n_v}$	Set of bivariate polynomials of relative degrees n_h and n_v (with respect to the indeterminates $w_h \in \Im$ and $w_v \in \Im$, respectively) over $\Re$.
$\Re(w_h)_{n_h}(w_v)_{n_v}$	Set of rational bivariate polynomials of relative degrees n_h and n_v (with respect to the indeterminates $w_h \in \Im$ and $w_v \in \Im$, respectively) over $\Re$.
I_n	Unit matrix of size $n \times n$.
$\{a_{ij}\}$	Elements of matrix A.
A^*, A^T	Complex conjugate transpose and transpose of matrix A.
trace$[A]$, $\lambda_i[A]$	Trace and i-th eigenvalue of matrix A.
$\oplus$, $\otimes$	Matrix Kronecker sum and product operators.
$\mathbf{e}_i^{(n)}$	Unit vector in $\Re^n$ with 1 on the i-th row.
$E_{i,j}^{q\times p}$	$\mathbf{e}_i^{(q)}\mathbf{e}_j^{(p)^*} \in \Re^{q\times p}$.
$\overline{U}_{q\times p}$	$\sum_{i=1}^{q}\sum_{j=1}^{p} E_{i,j}^{(q\times p)} \otimes E_{i,j}^{(q\times p)} \in \Re^{q^2\times p^2}$.
$\|A\|_F$	Fröbenius norm of A.

For q-systems, indeterminate z (with or without a subscript) is used; for δ-systems, we use c (with or without a subscript). In the 1-D case, corresponding q- and δ-systems are related by

$$\delta = \frac{q-1}{\Delta} \Longleftrightarrow c = \frac{z-1}{\Delta},$$

where Δ is a positive real constant, usually the sampling time.

For 2-D systems, subscripts h and v denote horizontally propagating (h.p.) and vertically propagating (v.p.) subsystems of the corresponding Roesser local s.s. models.

n_h, n_v, n	Sizes of the h.p. and v.p. subsystems; $n = n_h + n_v$.
Δ_h, Δ_v	Positive real constants denoting 'sampling times' along h.p. and v.p. directions.
ξ, Ξ	$\Delta_h I_{n_h} \oplus \Delta_v I_{n_v} \in \Re^{n\times n}$, $\Delta_h I_{n_h q} \oplus \Delta_v I_{n_v q} \in \Re^{nq\times nq}$.
I_z, I_c	$z_h I_{n_h} \oplus z_v I_{n_v} \in \Im^{n\times n}$, $c_h I_{n_h} \oplus c_v I_{n_v} \in \Im^{n\times n}$.

Corresponding 2-D q- and δ-systems are related by

$$\delta_h = \frac{q_h - 1}{\Delta_h} \Longleftrightarrow c_h = \frac{z_h - 1}{\Delta_h}; \qquad \delta_v = \frac{q_v - 1}{\Delta_v} \Longleftrightarrow c_v = \frac{z_v - 1}{\Delta_v}.$$

We use subscripts δ and q to differentiate between corresponding δ- and q-systems; for example, s.s. realization of a given DT system is either $\{A_\delta, B_\delta, C_\delta, D_\delta\}$ if implemented based on δ-operator or $\{A_q, B_q, C_q, D_q\}$ if implemented based on q-operator. The following notation is also used:

$$H(c_h, c_v)|_{\mathbf{c}\to\mathbf{z}} = H(c_h, c_v)|_{\substack{c_h=(z_h-1)/\Delta_h \\ c_v=(z_v-1)/\Delta_v}}; \qquad G(z_h, z_v)|_{\mathbf{z}\to\mathbf{c}} = G(z_h, z_v)|_{\substack{z_h=1+\Delta_h c_h \\ z_v=1+\Delta_v c_v}}.$$

Stability studies of 2-D q- and δ-systems involve the following regions:

$\mathcal{U}_q^2, \overline{\mathcal{U}}_q^2, \mathcal{T}_q^2$ $\{(z_h, z_v) \in \Im^2 : |z_h| < 1, |z_v| < 1\}$, $\{(z_h, z_v) \in \Im^2 : |z_h| \leq 1, |z_v| \leq 1\}$, $\{(z_h, z_v) \in \Im^2 : |z_h| = 1, |z_v| = 1\}$.

$\mathcal{U}_\delta^2, \overline{\mathcal{U}}_\delta^2, \mathcal{T}_\delta^2$ $\{(c_h, c_v) \in \Im^2 : |c_h + 1/\Delta_h| < 1/\Delta_h, |c_v + 1/\Delta_v| < 1\}$, $\{(c_h, c_v) \in \Im^2 : |c_h + 1/\Delta_h| \leq 1/\Delta_h, |c_v + 1/\Delta_v| \leq 1\}$, $\{(c_h, c_v) \in \Im^2 : |c_h + 1/\Delta_h| = 1/\Delta_h, |c_v + 1/\Delta_v| = 1\}$.

A q-system polynomial with all its roots in $\mathcal{U}_q$ (for the 1-D case) or $\mathcal{U}_q^2$ (for the 2-D case) is said to be *stable*. The corresponding regions for a δ-system polynomial are $\mathcal{U}_\delta$ (for the 1-D case) and $\mathcal{U}_\delta^2$ (for the 2-D case), respectively.

2.2. Preliminaries

First, we provide a brief introduction to the Roesser local s.s. model applicable to 2-D q-operator based DT systems [8].

DEFINITION 2.1. The following partial ordering in $\aleph^2$ is used:

$$\begin{aligned}
(h,k) \leq (i,j) &\iff h \leq i \quad \text{and} \quad k \leq j; \\
(h,k) = (i,j) &\iff h = i \quad \text{and} \quad k = j; \\
(h,k) < (i,j) &\iff (h,k) \leq (i,j) \quad \text{and} \quad (h,k) \neq (i,j).
\end{aligned}$$

The 2-D dynamical system under consideration is assumed to be linear, shift-invariant, and strictly causal. Moreover, it is taken to be modeled by a set of first-order vector difference equations over $\Re$. Given such a p-input and q-output 2-D system, its n_hh-n_vv Roesser local s.s. model is of the form [8]

$$\begin{aligned}
\begin{bmatrix} q_h[\mathbf{x}^h](i,j) \\ q_v[\mathbf{x}^v](i,j) \end{bmatrix} &= \begin{bmatrix} A_q^{(1)} & A_q^{(2)} \\ A_q^{(3)} & A_q^{(4)} \end{bmatrix} \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} + \begin{bmatrix} B_q^{(1)} \\ B_q^{(2)} \end{bmatrix} \mathbf{u}(i,j) \doteq [A_q] \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} + [B_q]\mathbf{u}(i,j); \\
\mathbf{y}(i,j) &= \begin{bmatrix} C_q^{(1)} & C_q^{(2)} \end{bmatrix} \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} + [D_q]\mathbf{u}(i,j) \doteq [C_q] \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} + [D_q]\mathbf{u}(i,j),
\end{aligned} \tag{2.1}$$

where $\mathbf{u} \in \Re^p$, $\mathbf{x}^h \in \Re^{n_h}$, $\mathbf{x}^v \in \Re^{n_v}$, and $\mathbf{y} \in \Re^q$. Also, $A_q^{(1)} \in \Re^{n_h \times n_h}$, $A_q^{(2)} \in \Re^{n_h \times n_v}$, $A_q^{(3)} \in \Re^{n_v \times n_h}$, $A_q^{(4)} \in \Re^{n_v \times n_v}$, $B_q^{(1)} \in \Re^{n_h \times p}$, $B_q^{(2)} \in \Re^{n_v \times p}$, $C_q^{(1)} \in \Re^{q \times n_h}$, $C_q^{(2)} \in \Re^{q \times n_v}$, $D_q \in \Re^{q \times p}$, and $(i,j) \in \aleph^2$. The operators $q_h[\cdot]$ and $q_v[\cdot]$ denote

$$q_h[\mathbf{x}](i,j) = \mathbf{x}(i+1,j) \quad \text{and} \quad q_v[\mathbf{x}](i,j) = \mathbf{x}(i,j+1). \tag{2.2}$$

The s.s. model in (2.1) is typically denoted by the quadruple $\{A_q, B_q, C_q, D_q\}$. The corresponding 2-D characteristic equation and the 2-D transfer function it realizes are

$$\begin{aligned}
\det[I_z - A_q] &= \det[(z_h I_{n_h} \oplus z_v I_{n_v}) - A_q] \in \Re[z_h]_{n_h}[z_v]_{n_v}; \\
H_q(z_h, z_v) &= C_q(I_z - A_q)^{-1}B_q + D_q \in \Re(z_h)_{n_h}(z_v)_{n_v},
\end{aligned} \tag{2.3}$$

where $z_h, z_v \in \Im$. In the literature, $\mathbf{x}^h$ and $\mathbf{x}^v$ are referred to as the *horizontally propagating (h.p.)* and *vertically propagating (v.p.)* local state vectors of $\{A_q, B_q, C_q, D_q\}$.

Assuming no nonessential singularities of the second kind on $\mathcal{T}_q^2$, for BIBO stability of the s.s. model above, it is necessary and sufficient that (see [9], and references therein)

$$\det[I_z - A_q] \neq 0, \ \forall (z_h, z_v) \in \overline{U}_q^2. \tag{2.4}$$

For investigating coefficient sensitivity properties, we will use certain relationships encountered in Kronecker products and matrix differentiation. The following are from [10].

The derivative of $A = \{a_{ij}\} \in \Re^{q \times p}$ with respect to $b \in \Re$ is

$$\frac{\partial A}{\partial b} = \left\{ \frac{\partial a_{ij}}{\partial b} \right\} \in \Re^{q \times p}. \tag{2.5}$$

Hence

$$\left\| \frac{\partial A}{\partial b} \right\|_F^2 = \sum_{i=1}^{q} \sum_{j=1}^{p} \left(\frac{\partial a_{ij}}{\partial b} \right)^2. \tag{2.6}$$

The derivative of $A = \{a_{ij}\} \in \Re^{q \times p}$ with respect to $B = \{b_{k\ell}\} \in \Re^{s \times r}$ is the partitioned matrix whose (k, ℓ)-th partition is $\partial A / \partial b_{k\ell}$, that is,

$$\frac{\partial A}{\partial B} = \begin{bmatrix} \frac{\partial A}{\partial b_{11}} & \cdots & \frac{\partial A}{\partial b_{1r}} \\ \vdots & \ddots & \vdots \\ \frac{\partial A}{\partial b_{s1}} & \cdots & \frac{\partial A}{\partial b_{sr}} \end{bmatrix} \in \Re^{qs \times pr}. \tag{2.7}$$

Hence

$$\left\| \frac{\partial A}{\partial B} \right\|_F^2 = \sum_{i=1}^{q} \sum_{j=1}^{p} \sum_{k=1}^{s} \sum_{\ell=1}^{r} \left(\frac{\partial a_{ij}}{\partial b_{k\ell}} \right)^2 = \sum_{k=1}^{s} \sum_{\ell=1}^{r} \| \frac{\partial A}{\partial b_{k\ell}} \|_F^2. \tag{2.8}$$

III. State-Space Model for δ-Operator Implementation

3.1. Local s.s. model

To exploit the superior finite wordlength properties of δ-operator implementations, analogous to the 1-D case, let us define the operators $\delta_h[\cdot]$ and $\delta_v[\cdot]$ as follows:

$$\begin{aligned}\delta_h[\mathbf{x}](i,j) &= \frac{\mathbf{x}(i+1,j)-\mathbf{x}(i,j)}{\Delta_h} = \frac{q_h[\mathbf{x}](i,j)-\mathbf{x}(i,j)}{\Delta_h};\\ \delta_v[\mathbf{x}](i,j) &= \frac{\mathbf{x}(i,j+1)-\mathbf{x}(i,j)}{\Delta_v} = \frac{q_v[\mathbf{x}](i,j)-\mathbf{x}(i,j)}{\Delta_v},\end{aligned} \tag{3.1}$$

where Δ_h and Δ_v are two positive real numbers. Hence, the following relationships are applicable:

$$\delta_h = \frac{q_h-1}{\Delta_h} \iff q_h = 1+\Delta_h\delta_h; \qquad \delta_v = \frac{q_v-1}{\Delta_v} \iff q_v = 1+\Delta_v\delta_v. \tag{3.2}$$

Remark. When Δ_h and Δ_v are the 'sampling times' corresponding to the horizontal and vertical spatial directions, the operators δ_h and δ_v in fact provide the first-order forward Euler approximants of the corresponding derivatives. When $\Delta_h \to 0$ and $\Delta_v \to 0$, the operators δ_h and δ_v yield these derivatives. In the 1-D case, this is the reason for the possibility of a unified treatment of both CT and DT systems [5].

With (3.2) in mind, we get

$$\begin{aligned}\begin{bmatrix}\delta_h[\mathbf{x}^h](i,j)\\ \delta_v[\mathbf{x}^v](i,j)\end{bmatrix} &= \xi^{-1}\begin{bmatrix}(q_h-1)I_{n_h} & 0\\ 0 & (q_v-1)I_{n_v}\end{bmatrix}\begin{bmatrix}\mathbf{x}^h(i,j)\\ \mathbf{x}^v(i,j)\end{bmatrix}\\ \iff \begin{bmatrix}q_h[\mathbf{x}^h](i,j)\\ q_v[\mathbf{x}^v](i,j)\end{bmatrix} &= I_n + \xi\begin{bmatrix}\delta_h I_{n_h} & 0\\ 0 & \delta_v I_{n_v}\end{bmatrix}\begin{bmatrix}\mathbf{x}^h(i,j)\\ \mathbf{x}^v(i,j)\end{bmatrix}.\end{aligned} \tag{3.3}$$

Here,

$$\xi = [\Delta_h I_{n_h} \oplus \Delta_v I_{n_v}] \in \Re^{n\times n}. \tag{3.4}$$

Using (3.3) in (2.1), it is easy to get the following:

$$\begin{aligned}\begin{bmatrix}\delta_h[\mathbf{x}^h](i,j)\\ \delta_v[\mathbf{x}^v](i,j)\end{bmatrix} &= \begin{bmatrix}A_\delta^{(1)} & A_\delta^{(2)}\\ A_\delta^{(3)} & A_\delta^{(4)}\end{bmatrix}\begin{bmatrix}\mathbf{x}^h(i,j)\\ \mathbf{x}^v(i,j)\end{bmatrix} + \begin{bmatrix}B_\delta^{(1)}\\ B_\delta^{(2)}\end{bmatrix}\mathbf{u}(i,j) \doteq [A_\delta]\begin{bmatrix}\mathbf{x}^h(i,j)\\ \mathbf{x}^v(i,j)\end{bmatrix} + [B_\delta]\mathbf{u}(i,j);\\ \mathbf{y}(i,j) &= \begin{bmatrix}C_\delta^{(1)} & C_\delta^{(2)}\end{bmatrix}\begin{bmatrix}\mathbf{x}^h(i,j)\\ \mathbf{x}^v(i,j)\end{bmatrix} + [D_\delta]\mathbf{u}(i,j) \doteq [C_\delta]\begin{bmatrix}\mathbf{x}^h(i,j)\\ \mathbf{x}^v(i,j)\end{bmatrix} + [D_\delta]\mathbf{u}(i,j).\end{aligned} \tag{3.5}$$

In addition, as opposed to its corresponding q-operator implementation, in a δ-operator implementation, one must perform the following computations:

$$\mathbf{x}^h(i+1,j) = \mathbf{x}^h(i,j) + \Delta_h\cdot\delta_h[\mathbf{x}^h](i,j); \qquad \mathbf{x}^v(i,j+1) = \mathbf{x}^v(i,j) + \Delta_v\cdot\delta_v[\mathbf{x}^v](i,j). \tag{3.6}$$

Here,

$$\begin{aligned}&A_\delta = \xi^{-1}(A_q - I_n) \iff A_q = I_n + \xi A_\delta; \qquad B_\delta = \xi^{-1}B_q \iff B_q = \xi B_\delta;\\ &C_\delta = C_q \iff C_q = C_\delta; \qquad D_\delta = D_q \iff D_q = D_\delta.\end{aligned} \tag{3.7}$$

The size of each submatrix in (3.5) is equal to the corresponding submatrix in (2.1). In the sequel, the s.s. realization $\{A_q, B_q, C_q, D_q\}$ in (2.1) will be referred to as the *q-model,* while the s.s. realization $\{A_\delta, B_\delta, C_\delta, D_\delta\}$ in (3.5) will be referred to as the *δ-model.*

3.2. Properties of the δ-model

The general response equation of the δ-model may be derived in a manner that is exactly analogous to that in [8]. Hence, in what follows, only the salient results are given, detailed derivations being omitted for the sake of brevity.

The general response of δ-model is given by

$$\begin{aligned}\begin{bmatrix}\mathbf{x}^h(i,j)\\ \mathbf{x}^v(i,j)\end{bmatrix} &= \sum_{k=0}^{j} A_\delta^{i,j-k}\begin{bmatrix}\mathbf{x}^h(0,k)\\ \mathbf{0}\end{bmatrix} + \sum_{h=0}^{i} A_\delta^{i-j,j}\begin{bmatrix}\mathbf{0}\\ \mathbf{x}^v(h,0)\end{bmatrix}\\ &\quad + \sum_{(0,0)\le(h,k)<(i,j)}\left(A_\delta^{i-h-1,j-k}\xi\begin{bmatrix}B_\delta^{(1)}\\ \mathbf{0}\end{bmatrix} + A_\delta^{i-h,j-k-1}\xi\begin{bmatrix}\mathbf{0}\\ B_\delta^{(2)}\end{bmatrix}\right)\mathbf{u}(h,k); \qquad (3.8)\\ \mathbf{y}(i,j) &= \begin{bmatrix}C_\delta^{(1)} & C_\delta^{(2)}\end{bmatrix}\begin{bmatrix}\mathbf{x}^h(i,j)\\ \mathbf{x}^v(i,j)\end{bmatrix} + [D_\delta]\mathbf{u}(i,j).\end{aligned}$$

Here, $A_\delta^{i,j}$ refers to the *transition matrix* of δ-model. With the partial ordering in $\aleph^2$ agreed upon previously (Definition 2.1), it may be recursively computed as follows:

$$A_\delta^{i,j} = \begin{cases}\begin{bmatrix}0 & 0\\ 0 & 0\end{bmatrix}, & \text{for } (i,j) < (0,0);\\ \begin{bmatrix}I_{n_h} & 0\\ 0 & I_{n_v}\end{bmatrix}, & \text{for } (i,j) = (0,0);\\ \begin{bmatrix}I_{n_h} & 0\\ 0 & 0\end{bmatrix} + \xi\begin{bmatrix}A_\delta^{(1)} & A_\delta^{(2)}\\ 0 & 0\end{bmatrix}, & \text{for } (i,j) = (1,0);\\ \begin{bmatrix}0 & 0\\ 0 & I_{n_v}\end{bmatrix} + \xi\begin{bmatrix}0 & 0\\ A_\delta^{(3)} & A_\delta^{(4)}\end{bmatrix}, & \text{for } (i,j) = (0,1);\\ A_\delta^{1,0}A_\delta^{i-1,j} + A_\delta^{0,1}A_\delta^{i,j-1}, & \text{elsewhere.}\end{cases} \qquad (3.9)$$

Remarks.

1. $A_\delta^{1,0} + A_\delta^{0,1} = I_n + \xi A_\delta \iff A_\delta = \xi^{-1}(A_\delta^{1,0} + A_\delta^{0,1} - I_n)$.
2. $A_\delta^{i,0} = (A_\delta^{1,0})^i$, $\forall i \ge 1$, and $A_\delta^{0,j} = (A_\delta^{0,1})^j$, $\forall j \ge 1$.

The 2-D δ-model's characteristic equation and transfer function, and their relationships to those of the corresponding q-model, are as follows:

$$\begin{aligned}\det[I_c - A_\delta] &= \det[(c_h I_{n_h} \oplus c_v I_{n_v}) - A_\delta] = \frac{1}{\det[\xi]}\det[I_z - A_q]|_{\mathbf{z}\to\mathbf{c}} \in \Re[c_h]_{n_h}[c_v]_{n_v};\\ H_\delta(c_h,c_v) &= C_\delta(I_c - A_\delta)^{-1}B_\delta + D_\delta = H_q(z_h,z_v)|_{\mathbf{z}\to\mathbf{c}} \in \Re(c_h)_{n_h}(c_v)_{n_v},\end{aligned} \qquad (3.10)$$

where

$$c_h = \frac{z_h - 1}{\Delta_h} \iff z_h = 1 + \Delta_h c_h; \qquad c_v = \frac{z_v - 1}{\Delta_v} \iff z_v = 1 + \Delta_v c_v. \qquad (3.11)$$

As for the q-model, it is easy to show that, 2-D equivalent transformations of the type

$$\begin{bmatrix}\tilde{\mathbf{x}}^h(i,j)\\ \tilde{\mathbf{x}}^v(i,j)\end{bmatrix} = \begin{bmatrix}T^{(1)} & 0\\ 0 & T^{(4)}\end{bmatrix}\begin{bmatrix}\mathbf{x}^h(i,j)\\ \mathbf{x}^v(i,j)\end{bmatrix} \doteq [T]\begin{bmatrix}\mathbf{x}^h(i,j)\\ \mathbf{x}^v(i,j)\end{bmatrix}, \qquad (3.12)$$

where $T^{(1)} \in \Re^{n_h \times n_h}$ and $T^{(4)} \in \Re^{n_v \times n_v}$ are nonsingular, yield an equivalent 2-D s.s. realization $\{\tilde{A}_\delta, \tilde{B}_\delta, \tilde{C}_\delta, \tilde{D}_\delta\}$, where

$$\tilde{A}_\delta = TA_\delta T^{-1}, \quad \tilde{B}_\delta = TB_\delta, \quad \tilde{C}_\delta = C_\delta T^{-1}, \quad \text{and} \quad \tilde{D}_\delta = D_\delta. \qquad (3.13)$$

The transfer function of $\{\tilde{A}_\delta, \tilde{B}_\delta, \tilde{C}_\delta, \tilde{D}_\delta\}$ is the same as that for $\{A_\delta, B_\delta, C_\delta, D_\delta\}$.

We will also assume that

$$\det[I_c - A_\delta] \neq 0, \ \forall (c_h, c_v) \in \overline{U}_\delta^2. \tag{3.14}$$

Due to (2.4) and (3.14), assuming no nonessential singularities of the second kind on $\mathcal{T}_\delta^2$, this implies BIBO stability of the 2-D δ-model (see [11], and references therein).

3.3. Gramians

In the 2-D q-operator case, reachability and observability gramians are typically taken to be natural extensions of the integral expressions of their 1-D counterparts (see [12-13], and references therein). To adopt a similar approach for the δ-operator case, we first investigate 1-D gramians for the δ-operator case as defined in [5].

1-D case. We quote relevant definitions from [5], p. 194 and 200:

DEFINITION 3.1. [5]. Consider the 1-D stable δ-system $\{A_\delta, B_\delta, C_\delta, D_\delta\}$. The reachability gramian P_δ and observability gramian Q_δ are the unique solutions of the following Lyapunov equations:

$$A_\delta P_\delta + P_\delta A_\delta^* + \Delta \cdot A_\delta P_\delta A_\delta^* = -B_\delta B_\delta^*; \qquad A_\delta^* Q_\delta + Q_\delta A_\delta + \Delta \cdot A_\delta^* Q_\delta A_\delta = -C_\delta^* C_\delta.$$

We now provide the integral representations of P_δ and Q_δ:

LEMMA 3.1. Consider the 1-D stable δ-system $\{A_\delta, B_\delta, C_\delta, D_\delta\}$ with gramians P_δ and Q_δ. Let $\{A_q, B_q, C_q, D_q\}$ with gramians P_q and Q_q be the corresponding 1-D stable q-system. Then

$$P_\delta = \frac{1}{2\pi j} \oint_{\mathcal{T}_\delta} F_\delta(c) F_\delta^*(c) \frac{dc}{1 + \Delta c}; \qquad Q_\delta = \frac{1}{2\pi j} \oint_{\mathcal{T}_\delta} G_\delta^*(c) G_\delta(c) \frac{dc}{1 + \Delta c}.$$

Moreover,

$$P_\delta = \frac{1}{\Delta} P_q \iff P_q = \Delta P_\delta; \qquad Q_\delta = \Delta Q_q \iff Q_q = \frac{1}{\Delta} Q_\delta.$$

Proof. Note that, $A_q = I_n + \Delta A_\delta$, $B_q = \Delta B_\delta$, $C_q = C_\delta$, and $D_q = D_\delta$ [5]. Substitute these in the Lyapunov equation for P_δ in Definition 3.1 to get

$$A_q^* P_\delta A_q^* - P_\delta = -\frac{1}{\Delta} B_q B_q^*.$$

Noting that P_q is the unique solution of $A_q^* P_q A_q^* - P_q = -B_q B_q^*$, we have $P_\delta = P_q/\Delta$. Moreover, the integral expression for P_q is

$$P_q = \frac{1}{2\pi j} \oint_{\mathcal{T}_q} F_q(z) F_q^*(z) \frac{dz}{z},$$

where $F_q(z) \doteq (zI_n - A_q)^{-1} B_q$ [13]. The claim regarding P_δ now follows. The rest follows in a similar manner. ∎

2-D case. With Lemma 3.1 in mind, we now present the following

DEFINITION 3.2. Consider the 2-D stable δ-system $\{A_\delta, B_\delta, C_\delta, D_\delta\}$. The reachability gramian P_δ and observability gramian Q_δ are defined as

$$P_\delta \doteq \begin{bmatrix} P_\delta^{(1)} & P_\delta^{(2)} \\ P_\delta^{(3)} & P_\delta^{(4)} \end{bmatrix} = \frac{1}{(2\pi j)^2} \oint_{T_\delta^2} F_\delta(c_h, c_v) F_\delta^*(c_h, c_v) \frac{dc_h}{1+\Delta_h c_h} \frac{dc_v}{1+\Delta_v c_v};$$

$$Q_\delta \doteq \begin{bmatrix} Q_\delta^{(1)} & Q_\delta^{(2)} \\ Q_\delta^{(3)} & Q_\delta^{(4)} \end{bmatrix} = \frac{1}{(2\pi j)^2} \oint_{T_\delta^2} G_\delta^*(c_h, c_v) G_\delta(c_h, c_v) \frac{dc_h}{1+\Delta_h c_h} \frac{dc_v}{1+\Delta_v c_v},$$

where

$$F_\delta(c_h, c_v) \doteq (I_c - A_\delta)^{-1} B_\delta = \begin{bmatrix} \mathbf{f}_{\delta_1}^* \\ \mathbf{f}_{\delta_2}^* \\ \vdots \\ \mathbf{f}_{\delta_n}^* \end{bmatrix} \in \Re^{n\times p}(c_h)_{n_h}(c_v)_{n_v};$$

$$G_\delta(c_h, c_v) \doteq C_\delta(I_c - A_\delta)^{-1} = [\, \mathbf{g}_{\delta_1} \quad \mathbf{g}_{\delta_2} \quad \cdots \quad \mathbf{g}_{\delta_n} \,] \in \Re^{q\times n}(c_h)_{n_h}(c_v)_{n_v}.$$

Remarks.

1. Note that, $\mathbf{f}_{\delta_i}(c_h, c_v) \in \Im^p$, $\forall i = 1, \ldots, n$, and $\mathbf{g}_{\delta_j}(c_h, c_v) \in \Im^q$, $\forall j = 1, \ldots, n$.
2. To eventually compare the performance of the δ-model and its corresponding q-model, the following relationships will be useful:

$$\begin{aligned} &(I_c - A_\delta)^{-1}|_{\mathbf{c}\to\mathbf{z}} = (I_z - A_q)^{-1}\xi; \\ &F_\delta(c_h, c_v)|_{\mathbf{c}\to\mathbf{z}} = F_q(z_h, z_v) \Longleftrightarrow \mathbf{f}_{\delta_j}(c_h, c_v)|_{\mathbf{c}\to\mathbf{z}} = \mathbf{f}_{q_j}, \quad \text{for} \quad j = 1, \ldots, n; \\ &G_\delta(c_h, c_v)|_{\mathbf{c}\to\mathbf{z}} = G_q(z_h, z_v)\cdot\xi \Longleftrightarrow \mathbf{g}_{\delta_i}(c_h, c_v)|_{\mathbf{c}\to\mathbf{z}} = \begin{cases} \Delta_h \mathbf{g}_{q_i}, & \text{for } i = 1, \ldots, n_h; \\ \Delta_v \mathbf{g}_{q_i}, & \text{for } i = n_h+1, \ldots, n. \end{cases} \end{aligned} \tag{3.15}$$

3. Definition 3.2 is completely analogous to the 1-D and 2-D q-operator cases. In the latter case, these gramians have been extremely useful in, and hence, have been extensively used for, investigating coefficient sensitivity, roundoff noise propagation, model reduction, etc. For instance, see [12-16], and references therein.

LEMMA 3.2. Consider the 2-D stable δ-system $\{A_\delta, B_\delta, C_\delta, D_\delta\}$ with gramians P_δ and Q_δ. Let $\{A_q, B_q, C_q, D_q\}$ with gramians P_q and Q_q be the corresponding 2-D stable q-system. Then

$$P_\delta = \frac{1}{\Delta_h\Delta_v} P_q \iff P_q = \Delta_h\Delta_v P_\delta; \qquad Q_\delta = \frac{1}{\Delta_h\Delta_v} \xi Q_q \xi \iff Q_q = \Delta_h\Delta_v \xi^{-1} Q_\delta \xi^{-1}.$$

Proof. Consider the integral expression for P_δ in Definition 3.2. With the variable change $\mathbf{c} \to \mathbf{z}$ and (3.15), we get

$$P_\delta = \frac{1}{\Delta_h\Delta_v} \frac{1}{(2\pi j)^2} \oint \oint_{T_q^2} F_q(z_h, z_v) F_q^*(z_h, z_v) \frac{dz_h}{z_h} \frac{dz_v}{z_v}.$$

However [12],

$$P_q = \frac{1}{(2\pi j)^2} \oint \oint_{T_q^2} F_q(z_h, z_v) F_q^*(z_h, z_v) \frac{dz_h}{z_h} \frac{dz_v}{z_v}.$$

Hence, the claim regarding P_δ follows. The proof regarding Q_δ is similar. ■

COROLLARY 3.3. The block matrices of the gramians are related as follows:

$$\begin{bmatrix} P_\delta^{(1)} & P_\delta^{(2)} \\ P_\delta^{(3)} & P_\delta^{(4)} \end{bmatrix} = \frac{1}{\Delta_h\Delta_v} \begin{bmatrix} P_q^{(1)} & P_q^{(2)} \\ P_q^{(3)} & P_q^{(4)} \end{bmatrix} \iff \begin{bmatrix} P_q^{(1)} & P_q^{(2)} \\ P_q^{(3)} & P_q^{(4)} \end{bmatrix} = \Delta_h\Delta_v \begin{bmatrix} P_\delta^{(1)} & P_\delta^{(2)} \\ P_\delta^{(3)} & P_\delta^{(4)} \end{bmatrix};$$

$$\begin{bmatrix} Q_\delta^{(1)} & Q_\delta^{(2)} \\ Q_\delta^{(3)} & Q_\delta^{(4)} \end{bmatrix} = \begin{bmatrix} \frac{\Delta_h}{\Delta_v} Q_q^{(1)} & Q_q^{(2)} \\ Q_q^{(3)} & \frac{\Delta_v}{\Delta_h} Q_q^{(4)} \end{bmatrix} \iff \begin{bmatrix} Q_q^{(1)} & Q_q^{(2)} \\ Q_q^{(3)} & Q_q^{(4)} \end{bmatrix} = \begin{bmatrix} \frac{\Delta_v}{\Delta_h} Q_\delta^{(1)} & Q_\delta^{(2)} \\ Q_\delta^{(3)} & \frac{\Delta_h}{\Delta_v} Q_\delta^{(4)} \end{bmatrix}.$$

Proof. This follows directly from Lemma 3.2. ∎

With the above results in mind, we now make some pertinent statements that are in complete analogy with the 2-D q-operator case. These may be easily verified/justified from the corresponding results for the latter (see [12], and references therein).

LEMMA 3.4. The gramians may be represented as follows:

$$P_\delta = \frac{1}{\Delta_h \Delta_v} \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} M_{\delta_{ij}} M_{\delta_{ij}}^*; \qquad Q_\delta = \frac{1}{\Delta_h \Delta_v} \xi \cdot \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} A_\delta^{i,j^*} C_\delta^* C_\delta A_\delta^{i,j} \cdot \xi,$$

where

$$M_{\delta_{ij}} = \begin{cases} \mathbf{0}, & \text{for } (i,j) = (0,0); \\ A_\delta^{i-1,j} \xi \begin{bmatrix} B_\delta^{(1)} \\ \mathbf{0} \end{bmatrix} + A_\delta^{i,j-1} \xi \begin{bmatrix} \mathbf{0} \\ B_\delta^{(2)} \end{bmatrix}, & \text{for } (i,j) > (0,0). \end{cases}$$

LEMMA 3.5. Consider the 2-D stable δ-model $\{A_\delta, B_\delta, C_\delta, D_\delta\}$ with gramians P_δ and Q_δ. Let $\{\tilde{A}_\delta, \tilde{B}_\delta, \tilde{C}_\delta, \tilde{D}_\delta\}$ with gramians $\tilde{P}_\delta$ and $\tilde{Q}_\delta$ be an equivalent system obtained with a nonsingular transformation of the type in (3.12-13). Then, $\tilde{P}_\delta = T P_\delta T^*$ and $\tilde{Q}_\delta = T^{-1^*} Q_\delta T^{-1}$. Moreover, eigenvalues of PQ are invariant under such a transformation.

DEFINITION 3.3. The 2-D δ-model $\{A_\delta, B_\delta, C_\delta, D_\delta\}$ is said to be *balanced (BL)* if its gramians P_δ and Q_δ satisfy

$$P_\delta^{(1)} = Q_\delta^{(1)} \doteq \Sigma_\delta^{(1)} = \text{diag}\{\sigma_{\delta_1}^{(1)}, \sigma_{\delta_2}^{(1)}, \ldots, \sigma_{\delta_{n_h}}^{(1)}\}; \qquad P_\delta^{(4)} = Q_\delta^{(4)} \doteq \Sigma_\delta^{(4)} = \text{diag}\{\sigma_{\delta_1}^{(4)}, \sigma_{\delta_2}^{(4)}, \ldots, \sigma_{\delta_{n_v}}^{(4)}\}.$$

We refer to $\sigma_{\delta_i}^{(1)}$, $i = 1, \ldots, n_h$, and $\sigma_{\delta_j}^{(4)}$, $j = 1, \ldots, n_v$, as the *Hankel singular values* of h.p. and v.p. subsystems, repectively.

If the principal block diagonal matrices of P_δ and Q_δ are each positive definite, a corresponding BL realization may be obtained through a certain simultaneous diagonalization procedure referred to as *Laub's algorithm* [17]. Regarding this, we have

LEMMA 3.6. Local reachability and observability of δ-model $\{A_\delta, B_\delta, C_\delta, D_\delta\}$ and its corresponding q-model $\{A_q, B_q, C_q, D_q\}$ are equivalent. Moreover, when $\{A_\delta, B_\delta, C_\delta, D_\delta\}$ is locally reachable and observable, $P_\delta^{(1)}$, $P_\delta^{(4)}$, $Q_\delta^{(1)}$, and $Q_\delta^{(4)}$ are each positive definite.

Separable systems. A separable (in denominator) 2-D q-system has the property that $A_q^{(2)} = \mathbf{0}$ (or, equivalently, $A_q^{(3)} = \mathbf{0}$). For such a system, off-diagonal submatrices of P_q and Q_q are all zero [12]. Moreover, the diagonal submatrices may be conveniently computed through the solution of two pairs of Lyapunov equations.

From (3.7), it is clear that, a separable 2-D q-system gives rise to a separable 2-D δ-system. Regarding the corresponding gramians, we may state the following

THEOREM 3.7. Consider the separable 2-D δ-system $\{A_\delta, B_\delta, C_\delta, D_\delta\}$ with gramians P_δ and Q_δ. Then, $P_\delta^{(2)} = Q_\delta^{(2)} = \mathbf{0}$ and $P_\delta^{(3)} = Q_\delta^{(3)} = \mathbf{0}$. Moreover, the diagonal block matrices of P_δ and Q_δ may be

computed through solution of the following two pairs of Lyapunov equations:

$$A_\delta^{(1)} P_\delta^{(1)} + P_\delta^{(1)} A_\delta^{(1)^*} + \Delta_h A_\delta^{(1)} P_\delta^{(1)} A_\delta^{(1)^*} = -\frac{1}{\Delta_v} B_\delta^{(1)} B_\delta^{(1)^*};$$

$$A_\delta^{(1)^*} Q_\delta^{(1)} + Q_\delta^{(1)} A_\delta^{(1)} + \Delta_h A_\delta^{(1)^*} Q_\delta^{(1)} A_\delta^{(1)} = -\frac{1}{\Delta_v} \begin{bmatrix} C_\delta^{(1)} & R_\delta^{(4)} A_\delta^{(3)} \end{bmatrix}^* \begin{bmatrix} C_\delta^{(1)} & R_\delta^{(4)} A_\delta^{(3)} \end{bmatrix};$$

$$A_\delta^{(4)} P_\delta^{(4)} + P_\delta^{(4)} A_\delta^{(4)^*} + \Delta_v A_\delta^{(4)} P_\delta^{(4)} A_\delta^{(4)^*} = -\frac{1}{\Delta_h} \begin{bmatrix} B_\delta^{(2)} & A_\delta^{(3)} S_\delta^{(1)} \end{bmatrix} \begin{bmatrix} B_\delta^{(2)} & A_\delta^{(3)} S_\delta^{(1)} \end{bmatrix}^*;$$

$$A_\delta^{(4)^*} Q_\delta^{(4)} + Q_\delta^{(4)} A_\delta^{(4)} + \Delta_v A_\delta^{(4)^*} Q_\delta^{(4)} A_\delta^{(4)} = -\frac{1}{\Delta_h} C_\delta^{(2)^*} C_\delta^{(2)},$$

where $R_\delta^{(4)^*} R_\delta^{(4)} = \Delta_h \Delta_v Q_\delta^{(4)}$ and $S_\delta^{(1)} S_\delta^{(1)^*} = \Delta_h \Delta_v P_\delta^{(1)}$.

Proof. Results regarding the off-diagonal submatrices are obvious from Corollary 3.3. Regarding the diagonal submatrices, claim may be shown using Theorem 3.2.2 of [12]. For instance, consider the Lypaunov equation

$$A_q^{(1)^*} Q_q^{(1)} A^{(1)} - Q_q^{(1)} = -C_q^{(1)^*} C_q^{(1)} - A_q^{(3)^*} Q_q^{(4)} A_q^{(3)}.$$

Using (3.7) and Corollary 3.3, second Lyapunov equation in the claim results. Rest follows in a similar manner. ■

3.4. Computation of BL Realizations

Computation of gramians and obtaining BL realizations for q-systems have been investigated quite thoroughly. In the 1-D and 2-D separable cases, one may solve Lyapunov equations and use Laub's algorithm [17]. In the 2-D non-separable case, this computation is not that easy; however, several techniques have been developed [12], [18].

In this section, we provide the relationship between BL realizations of corresponding δ- and q-models. This allows all available techniques for gramian computation of q-systems to be utilized for δ-systems as well. We believe this to be an important contribution, and, to the authors' knowledge, such a relationship is not available even for the 1-D case. The development below concentrates on the 2-D case; the 1-D case is even simpler. For convenience, we use the following notation:
$\{A, B, C, D\} \xrightarrow{T} \{\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}\}$: This denotes $\tilde{A} = TAT^{-1}$, $\tilde{B} = TB$, $\tilde{C} = CT^{-1}$, and $\tilde{D} = D$, where T is of type (3.12-13).
$\{A_q, B_q, C_q, D_q\} \xrightarrow{q \to \delta} \{A_\delta, B_\delta, C_\delta, D_\delta\}$: This denotes the corresponding δ-system obtained by applying (3.7).
$\{A_\delta, B_\delta, C_\delta, D_\delta\} \xrightarrow{\delta \to q} \{A_q, B_q, C_q, D_q\}$: This denotes the corresponding q-system obtained by applying (3.7).

Moreover, we use the following:

$\{A_q, B_q, C_q, D_q\}$	Given 2-D q-system.
$\{A_{qB}, B_{qB}, C_{qB}, D_{qB}\}$	BL realization of $\{A_q, B_q, C_q, D_q\}$ obtained by applying T_q, that is, $\{A_q, B_q, C_q, D_q\} \xrightarrow{T_q} \{A_{qB}, B_{qB}, C_{qB}, D_{qB}\}$.
$\{A_\delta, B_\delta, C_\delta, D_\delta\}$	2-D δ-system obtained by applying (3.7) to $\{A_q, B_q, C_q, D_q\}$, that is, $\{A_q, B_q, C_q, D_q\} \xrightarrow{q \to \delta} \{A_\delta, B_\delta, C_\delta, D_\delta\}$.
$\{A_{\delta B}, B_{\delta B}, C_{\delta B}, D_{\delta B}\}$	BL realization of $\{A_\delta, B_\delta, C_\delta, D_\delta\}$ obtained by applying T_δ, that is, $\{A_\delta, B_\delta, C_\delta, D_\delta\} \xrightarrow{T_\delta} \{A_{\delta B}, B_{\delta B}, C_{\delta B}, D_{\delta B}\}$.
$\{A_{\delta B2q}, B_{\delta B2q}, C_{\delta B2q}, D_{\delta B2q}\}$	2-D q-system obtained by applying (3.7) to $\{A_{\delta B}, B_{\delta B}, C_{\delta B}, D_{\delta B}\}$, that is, $\{A_{\delta B}, B_{\delta B}, C_{\delta B}, D_{\delta B}\} \xrightarrow{\delta \to q} \{A_{\delta B2q}, B_{\delta B2q}, C_{\delta B2q}, D_{\delta B2q}\}$.
$\{A_{qB2\delta}, B_{qB2\delta}, C_{qB2\delta}, D_{qB2\delta}\}$	2-D δ-system obtained by applying (3.7) to $\{A_{qB}, B_{qB}, C_{qB}, D_{qB}\}$, that is, $\{A_{qB}, B_{qB}, C_{qB}, D_{qB}\} \xrightarrow{q \to \delta} \{A_{qB2\delta}, B_{qB2\delta}, C_{qB2\delta}, D_{qB2\delta}\}$.

LEMMA 3.8. The following relationships are true:

$$\{A_q, B_q, C_q, D_q\} \xrightarrow{T_\delta} \{A_{\delta B2q}, B_{\delta B2q}, C_{\delta B2q}, D_{\delta B2q}\}; \qquad \{A_\delta, B_\delta, C_\delta, D_\delta\} \xrightarrow{T_q} \{A_{qB2\delta}, B_{qB2\delta}, C_{qB2\delta}, D_{qB2\delta}\}.$$

Proof. First,

$$A_{\delta B2q} = I_n + \xi A_{\delta B} = I_n + \xi T_\delta A_\delta T_\delta^{-1} = I_n + \xi T_\delta \xi^{-1}(A_q - I_n)T_\delta^{-1} = T_\delta A_q T_\delta^{-1},$$

since $\xi T_\delta \xi^{-1} = T_\delta$. The remainder may proven in a similar manner. ∎

LEMMA 3.9. The following relationships are true:

$$\{A_{\delta B2q}, B_{\delta B2q}, C_{\delta B2q}, D_{\delta B2q}\} \xrightarrow{\xi^{-1/2}} \{A_{qB}, B_{qB}, C_{qB}, D_{qB}\};$$
$$\{A_{qB2\delta}, B_{qB2\delta}, C_{qB2\delta}, D_{qB2\delta}\} \xrightarrow{\xi^{1/2}} \{A_{\delta B}, B_{\delta B}, C_{\delta B}, D_{\delta B}\}.$$

Proof. Note that, $\{A_{\delta B}, B_{\delta B}, C_{\delta B}, D_{\delta B}\}$ has following gramians:

$$P_{\delta B} = \begin{bmatrix} \Sigma_\delta^{(1)} & P_{\delta B}^{(2)} \\ P_{\delta B}^{(3)} & \Sigma_\delta^{(4)} \end{bmatrix}; \quad Q_{\delta B} = \begin{bmatrix} \Sigma_\delta^{(1)} & Q_{\delta B}^{(2)} \\ Q_{\delta B}^{(3)} & \Sigma_\delta^{(4)} \end{bmatrix}.$$

Hence, from Corollary 3.3, gramians of $\{A_{\delta B2q}, B_{\delta B2q}, C_{\delta B2q}, D_{\delta B2q}\}$ are as follows:

$$P_{\delta B2q} = \Delta_h \Delta_v \begin{bmatrix} \Sigma_\delta^{(1)} & P_{\delta B}^{(2)} \\ P_{\delta B}^{(3)} & \Sigma_\delta^{(4)} \end{bmatrix}; \quad Q_{\delta B2q} = \begin{bmatrix} \frac{\Delta_v}{\Delta_h}\Sigma_\delta^{(1)} & Q_{\delta B}^{(2)} \\ Q_{\delta B}^{(3)} & \frac{\Delta_h}{\Delta_v}\Sigma_\delta^{(4)} \end{bmatrix}.$$

To get $\{A_{qB}, B_{qB}, C_{qB}, D_{qB}\}$, we need to simultaneously diagonalize the two pairs $\{\Delta_h\Delta_v\Sigma_\delta^{(1)}, (\Delta_v/\Delta_h)\Sigma_\delta^{(1)}\}$ and $\{\Delta_h\Delta_v\Sigma_\delta^{(4)}, (\Delta_h/\Delta_v)\Sigma_\delta^{(4)}\}$. By applying Laub's algorithm, we get these two transformations to be $\Delta_h^{-1/2} I_{n_h}$ and $\Delta_v^{-1/2} I_{n_v}$. This proves the first part. The remainder follows in a similar manner. ∎

COROLLARY 3.10. The relationship between $\{A_{qB}, B_{qB}, C_{qB}, D_{qB}\}$ and $\{A_{\delta B}, B_{\delta B}, C_{\delta B}, D_{\delta B}\}$ is as follows:

$$A_{\delta B} = \xi^{-1/2}(A_{qB} - I_n)\xi^{-1/2}; \quad B_{\delta B} = \xi^{-1/2} B_{qB}; \quad C_{\delta B} = C_{qB}\xi^{-1/2}; \quad D_{\delta B} = D_{qB}.$$

Proof. Note that, from Lemma 3.9,

$$A_{\delta B} = \xi^{-1}(A_{\delta B2q} - I_n) = \xi^{-1}(\xi^{1/2} A_{qB} \xi^{-1/2} - I_n) = \xi^{-1/2}(A_{qB} - I_n)\xi^{-1/2}.$$

The rest follows in a similar manner. ∎

The above are summarized below. Note that, the missing 'links' may also be easily obtained.

$$\begin{array}{ccccc} \{A_q, B_q, C_q, D_q\} & \xrightarrow{T_q} & \{A_{qB}, B_{qB}, C_{qB}, D_{qB}\} & \xleftarrow{\xi^{-1/2}} & \{A_{\delta B2q}, B_{\delta B2q}, C_{\delta B2q}, D_{\delta B2q}\} \\ \downarrow{\scriptstyle q\to\delta} & & \downarrow \text{Corollary 3.10} & & \\ \{A_\delta, B_\delta, C_\delta, D_\delta\} & \xrightarrow{\xi^{1/2}T_q} & \{A_{\delta B}, B_{\delta B}, C_{\delta B}, D_{\delta B}\} & \xleftarrow{\xi^{1/2}} & \{A_{qB2\delta}, B_{qB2\delta}, C_{qB2\delta}, D_{qB2\delta}\} \end{array}$$

IV. Coefficient Sensitivity

Coefficient sensitivity is an important criterion on which one s.s. realization may be preferred over another. By generalizing a certain sensitivity measure in [19], Lutz and Hakimi [20] have addressed sensitivity minimization of MIMO 1-D CT systems. The 2-D q-operator case appears in [15], and references therein. This work, applicable only to the SISO case, reveals that realizations possessing minimum coefficient sensitivity are equivalent to BL (modulo a block orthogonal similarity transformation) realizations (see [12], and references therein).

In what follows, we study coefficient sensitivity properties of the 2-D δ-model introduced in Section III. Both FXP and FLP arithmetic implementations are addressed. We follow a more direct approach via Kronecker product formulation and, as a result, this work is applicable to the more general MIMO case.

In practice, effects of coefficient sensitivity appear in the system frequency response. Hence, it is appropriate to study the quantities $\partial H_\delta/\partial A_\delta$, $\partial H_\delta/\partial B_\delta$, $\partial H_\delta/\partial C_\delta$ and $\partial H_\delta/\partial D_\delta$. Using relationships regarding matrix Kronecker products taken from the excellent treatise of Brewer [10], we first develop certain relationships regarding these quantities. For the readers' convenience, relationships used from [10] are identified by the same equation numbers (these begin with the letter T).

First,

$$\begin{aligned} S_{\delta_{A_\delta}}(c_h, c_v) \doteq \frac{\partial}{\partial A_\delta} H_\delta(c_h, c_v) &= \frac{\partial}{\partial A_\delta}[C_\delta(I_c - A_\delta)^{-1}B_\delta + D_\delta] \\ &= [I_n \otimes C_\delta][I_n \otimes (I_c - A_\delta)^{-1}] \cdot \frac{\partial}{\partial A_\delta}[I_c - A_\delta] \cdot [I_n \otimes (I_c - A_\delta)^{-1}][I_n \otimes B_\delta] \\ &\qquad \text{from (T4.3) and (T5.5)} \\ &= [I_n \otimes C_\delta(I_c - A_\delta)^{-1}] \cdot \overline{U}_{n\times n} \cdot [I_n \otimes (I_c - A_\delta)^{-1}B_\delta] \\ &\qquad \text{from (T2.4) and (T5.1).} \end{aligned}$$

Hence

$$S_{\delta_{A_\delta}}(c_h, c_v) = [I_n \otimes G_\delta] \cdot \overline{U}_{n\times n} \cdot [I_n \otimes F_\delta] \in \Im^{nq\times np}. \tag{4.1}$$

Second,

$$\begin{aligned} S_{\delta_{B_\delta}}B(c_h, c_v) \doteq \frac{\partial}{\partial B_\delta} H_\delta(c_h, c_v) &= \frac{\partial}{\partial B_\delta}[C_\delta(I_c - A_\delta)^{-1}B_\delta + D_\delta] = \frac{\partial}{\partial B_\delta}[G_\delta B_\delta] \\ &= [I_n \otimes G_\delta] \cdot \frac{\partial B_\delta}{\partial B_\delta} \quad \text{from (T4.3).} \end{aligned}$$

Hence

$$S_{\delta_{B_\delta}}(c_h, c_v) = [I_n \otimes G_\delta] \cdot \overline{U}_{n\times p} \in \Im^{nq\times p^2}. \tag{4.2}$$

Third,

$$\begin{aligned} S_{\delta_{C_\delta}}(c_h, c_v) \doteq \frac{\partial}{\partial C_\delta} H_\delta(c_h, c_v) &= \frac{\partial}{\partial C_\delta}[C_\delta(I_c - A_\delta)^{-1}B_\delta + D_\delta] = \frac{\partial}{\partial C_\delta}[C_\delta F_\delta] \\ &= \frac{\partial C_\delta d}{\partial C_\delta} \cdot [I_n \otimes F_\delta] \quad \text{from (T4.3).} \end{aligned}$$

Hence

$$S_{\delta_{C_\delta}}(c_h, c_v) = \overline{U}_{q\times n} \cdot [I_n \otimes F_\delta] \in \Im^{q^2\times np}. \tag{4.3}$$

Fourth,

$$S_{\delta_{D_\delta}}(c_h, c_v) \doteq \frac{\partial}{\partial D_\delta} H_\delta(c_h, c_v) = \frac{\partial}{\partial D_\delta}[C_\delta(I_c - A_\delta)^{-1}B_\delta + D_\delta]$$
$$= \frac{\partial D_\delta}{\partial D_\delta}.$$

Hence

$$S_{\delta_{D_\delta}}(c_h, c_v) = \overline{U}_{q\times p} \in \Re^{q^2\times p^2}. \tag{4.4}$$

LEMMA 4.1 The quantities $S_{\delta_{A_\delta}}(c_h, c_v)$, $S_{\delta_{B_\delta}}(c_h, c_v)$, $S_{\delta_{C_\delta}}(c_h, c_v)$, and $S_{\delta_{D_\delta}}(c_h, c_v)$ of the δ-model are

$$S_{\delta_{A_\delta}}(c_h, c_v) = \begin{bmatrix} \mathbf{g}_{\delta_1} \\ \mathbf{g}_{\delta_2} \\ \vdots \\ \mathbf{g}_{\delta_n} \end{bmatrix} [\mathbf{f}^*_{\delta_1} \quad \mathbf{f}^*_{\delta_2} \quad \cdots \quad \mathbf{f}^*_{\delta_n}]; \qquad S_{\delta_{B_\delta}}(c_h, c_v) = \begin{bmatrix} \mathbf{g}^{(1)}_{\delta_1} & \mathbf{g}^{(2)}_{\delta_1} & \cdots & \mathbf{g}^{(p)}_{\delta_1} \\ \mathbf{g}^{(1)}_{\delta_2} & \mathbf{g}^{(2)}_{\delta_2} & \cdots & \mathbf{g}^{(p)}_{\delta_2} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{g}^{(1)}_{\delta_n} & \mathbf{g}^{(2)}_{\delta_n} & \cdots & \mathbf{g}^{(p)}_{\delta_n} \end{bmatrix};$$

$$S_{\delta_{C_\delta}}(c_h, c_v) = \begin{bmatrix} \mathbf{f}^{(1)^*}_{\delta_1} & \mathbf{f}^{(1)^*}_{\delta_2} & \cdots & \mathbf{f}^{(1)^*}_{\delta_n} \\ \mathbf{f}^{(2)^*}_{\delta_1} & \mathbf{f}^{(2)^*}_{\delta_2} & \cdots & \mathbf{f}^{(2)^*}_{\delta_n} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{f}^{(q)^*}_{\delta_1} & \mathbf{f}^{(q)^*}_{\delta_2} & \cdots & \mathbf{f}^{(q)^*}_{\delta_n} \end{bmatrix}; \qquad S_{\delta_{D_\delta}}(c_h, c_v) = \begin{bmatrix} E_{1,1} & E_{1,2} & \cdots & E_{1,p} \\ E_{2,1} & E_{2,2} & \cdots & E_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ E_{q,1} & E_{q,2} & \cdots & E_{q,p} \end{bmatrix}.$$

Here, $\mathbf{f}^{(j)^*}_{\delta_i}$ denotes a $q \times p$ null matrix except its j-th row which is $\mathbf{f}^*_{\delta_i}$, $\mathbf{g}^{(j)}_{\delta_i}$ denotes a $q \times p$ null matrix except its j-th column which is $\mathbf{g}_{\delta_i}$, and $E_{i,j}$ are $n \times p$ elementary matrices.

Proof. Relationship for $S_{\delta_{D_\delta}}$ follows immediately from (4.4). To show the remainder, note that

$$[I_n \otimes F_\delta] = \begin{bmatrix} F_\delta & 0 & \cdots & 0 \\ 0 & F_\delta & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & F_\delta \end{bmatrix} \in \Im^{n^2\times np}; \qquad [I_n \otimes G_\delta] = \begin{bmatrix} G_\delta & 0 & \cdots & 0 \\ 0 & G_\delta & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & G_\delta \end{bmatrix} \in \Im^{nq\times n^2}.$$

Here, $[I_n \otimes F_\delta]$ and $[I_n \otimes G_\delta]$ each has $n \times n$ blocks. Claim now follows through simple yet tedious algebraic manipulations. ∎

COROLLARY 4.2. The quantities $S_{\delta_{A_\delta}}(c_h, c_v)$, $S_{\delta_{B_\delta}}(c_h, c_v)$, $S_{\delta_{C_\delta}}(c_h, c_v)$, and $S_{\delta_{D_\delta}}(c_h, c_v)$ of δ-model and the quantities $S_{q_{A_q}}(z_h, z_v)$, $S_{q_{B_q}}(z_h, z_v)$, $S_{q_{C_q}}(z_h, z_v)$, and $S_{q_{D_q}}(z_h, z_v)$ of the corresponding q-model are related through the following:

$$S_{\delta_{A_\delta}}(c_h, c_v)|_{\mathbf{c}\to\mathbf{z}} = \Xi S_{q_{A_q}}(z_h, z_v); \qquad S_{\delta_{B_\delta}}(c_h, c_v)|_{\mathbf{c}\to\mathbf{z}} = \Xi S_{q_{B_q}}(z_h, z_v);$$
$$S_{\delta_{C_\delta}}(c_h, c_v)|_{\mathbf{c}\to\mathbf{z}} = S_{q_{C_q}}(z_h, z_v); \qquad S_{\delta_{D_\delta}}(c_h, c_v)|_{\mathbf{c}\to\mathbf{z}} = S_{q_{D_q}}(z_h, z_v),$$

where $\Xi \doteq [\Delta_h I_{n_h q} \oplus \Delta_v I_{n_v q}]$.

Proof. This is immediate when (3.15) is applied to Lemma 4.1. ∎

To proceed further, we utilize the following

DEFINITION 4.1. Let $H_\delta(c_h, c_v)$ be a bivariate matrix-valued function that is analytic on $\mathcal{T}^2_\delta$. Then,

$$\|H_\delta(c_h, c_v)\|_p \doteq \left[\frac{1}{(2\pi\jmath)^2} \oint \oint_{\mathcal{T}^2_q} \|H_\delta(c_h, c_v)|_{\mathbf{c}\to\mathbf{z}}\|^p_F \frac{dz_h}{z_h}\frac{dz_v}{z_v}\right]^{\frac{1}{p}}.$$

Remark. This matrix norm is extensively utilized in work related to coefficient sensitivity (see [15], and references therein) mainly because it leads to tractable results. This, and our desire to make a comparison with the corresponding q-model, are the primary reasons for its use here.

FXP Arithmetic Case

For FXP arithmetic implementations, it is appropriate to define an *absolute* sensitivity measure as

$$M_{\delta_{\mathrm{FXP}}} = \|S_{\delta_{A_\delta}}\|_1^2 + \frac{1}{p}\|S_{\delta_{B_\delta}}\|_2^2 + \frac{1}{q}\|S_{\delta_{C_\delta}}\|_2^2 + \frac{1}{pq}\|S_{\delta_{D_\delta}}\|_2^2, \tag{4.5}$$

where, as defined previously,

$$\begin{aligned} S_{\delta_{A_\delta}} &= \{s_{\delta_{A_\delta},ij}\} = \partial H_\delta/\partial a_{\delta_{ij}}; \qquad S_{\delta_{B_\delta}} = \{s_{\delta_{B_\delta},ij}\} = \partial H_\delta/\partial b_{\delta_{ij}}; \\ S_{\delta_{C_\delta}} &= \{s_{\delta_{C_\delta},ij}\} = \partial H_\delta/\partial c_{\delta_{ij}}; \qquad S_{\delta_{D_\delta}} = \{s_{\delta_{D_\delta},ij}\} = \partial H_\delta/\partial d_{\delta_{ij}}. \end{aligned} \tag{4.6}$$

Remarks.

1. $M_{\delta_{\mathrm{FXP}}}$ takes into account variations in frequency response with respect to perturbations in A_δ, B_δ, C_δ, and D_δ. Note that, in FXP, possible perturbation of a particular coefficient is approximately independent of its nominal value. This justifies the definition in (4.5).
2. Use of different norms is for mathematical feasibility and tractability, and is typical in coefficient sensitivity studies [15], [3]. Given a δ-model $\{A_\delta, B_\delta, C_\delta, D_\delta\}$, the objective is to characterize those realizations belonging to the class $\{\tilde{A}_\delta, \tilde{B}_\delta, \tilde{C}_\delta, \tilde{D}_\delta\} \equiv \{TA_\delta T^{-1}, TB_\delta, C_\delta T^{-1}, D_\delta\}$, where T is of the type in (3.12-13), that minimize $M_{\delta_{\mathrm{FXP}}}$.
3. Weights associated with each term in (4.5) may be thought of as *averaging factors.* The ensuing measure then may be thought of as an *average sensitivity per input/output.*
4. In a δ-operator implementation, due to the necessity of performing the computation in (3.6), coefficient sensitivity will be affected by perturbations of Δ_h and Δ_v as well. Hence, $M_{\delta_{\mathrm{FXP}}}$ must be modified to contain terms of the nature $\|S_{\delta_{\Delta_h}}\|_2$ and $\|S_{\delta_{\Delta_v}}\|_2$. However, selection of Δ_h and Δ_v is somewhat arbitrary, and they may be chosen to possess exact binary FXP representations. If so, corresponding sensitivity terms may be neglected. In what follows, we assume that this is the case.

We now attempt to obtain an expression for $M_{\delta_{\mathrm{FXP}}}$ as follows:

$$\begin{aligned} \|S_{\delta_{A_\delta}}\|_1^2 &= \left[\frac{1}{(2\pi\jmath)^2}\oint\oint_{T_q^2}\left\|\begin{bmatrix}\mathbf{g}_{\delta_1}\\ \vdots \\ \mathbf{g}_{\delta_n}\end{bmatrix}[\,\mathbf{f}^*_{\delta_1} \quad \cdots \quad \mathbf{f}^*_{\delta_n}\,]|_{\mathbf{c}\to\mathbf{z}}\right\|_F \frac{dz_h}{z_h}\frac{dz_v}{z_v}\right]^2 \\ &\le \left[\frac{1}{(2\pi\jmath)^2}\oint\oint_{T_q^2}\left\|\begin{bmatrix}\mathbf{g}_{\delta_1}\\ \vdots \\ \mathbf{g}_{\delta_n}\end{bmatrix}|_{\mathbf{c}\to\mathbf{z}}\right\|_F^2 \frac{dz_h}{z_h}\frac{dz_v}{z_v}\right] \\ &\quad \cdot\left[\frac{1}{(2\pi\jmath)^2}\oint\oint_{T_q^2}\|[\,\mathbf{f}^*_{\delta_1} \quad \cdots \quad \mathbf{f}^*_{\delta_n}\,]|_{\mathbf{c}\to\mathbf{z}}\|_F^2 \frac{dz_h}{z_h}\frac{dz_v}{z_v}\right] \\ &= \mathrm{trace}\left[\frac{1}{(2\pi\jmath)^2}\oint\oint_{T_q^2} G^*_\delta(c_h,c_v)G_\delta(c_h,c_v)|_{\mathbf{c}\to\mathbf{z}}\frac{dz_h}{z_h}\frac{dz_v}{z_v}\right] \\ &\quad \cdot \mathrm{trace}\left[\frac{1}{(2\pi\jmath)^2}\oint\oint_{T_q^2} F_\delta(c_h,c_v)F^*_\delta(c_h,c_v)|_{\mathbf{c}\to\mathbf{z}}\frac{dz_h}{z_h}\frac{dz_v}{z_v}\right]. \end{aligned}$$

To get the first inequality, we have used mutual consistency of Fröbenius norm, that is, $\|AB\|_F \leq \|A\|_F \cdot \|B\|_F$, and Cauchy-Schwarz inequality; last equality follows due to $\|A\|_F^2 = \text{trace}[A^*A]$ [21]. Hence, using (3.15),

$$\|S_{\delta_{A_\delta}}\|_1^2 \leq \text{trace}[P_q] \cdot \text{trace}[\xi Q_q \xi] = (\Delta_h \Delta_v)^2 \cdot \text{trace}[P_\delta] \cdot \text{trace}[Q_\delta]. \tag{4.7a}$$

Next,

$$\begin{aligned}
\|S_{\delta_{B_\delta}}\|_2^2 &= \left[\frac{1}{(2\pi\jmath)^2} \oint \oint_{\mathcal{T}_q^2} \left\| \begin{bmatrix} g_{\delta_1}^{(1)} & \cdots & g_{\delta_1}^{(p)} \\ \vdots & \ddots & \vdots \\ g_{\delta_n}^{(1)} & \cdots & g_{\delta_n}^{(p)} \end{bmatrix} |_{\mathbf{c}\to\mathbf{z}} \right\|_F^2 \frac{dz_h}{z_h} \frac{dz_v}{z_v} \right] \\
&= \left[\frac{1}{(2\pi\jmath)^2} \oint \oint_{\mathcal{T}_q^2} p \|G_\delta(c_h, c_v)|_{\mathbf{c}\to\mathbf{z}}\|_F^2 \frac{dz_h}{z_h} \frac{dz_v}{z_v} \right] \\
&= p \cdot \text{trace} \left[\frac{1}{(2\pi\jmath)^2} \oint \oint_{\mathcal{T}_q^2} G_\delta^*(c_h, c_v) G_\delta(c_h, c_v)|_{\mathbf{c}\to\mathbf{z}} \frac{dz_h}{z_h} \frac{dz_v}{z_v} \right]
\end{aligned}$$

Hence,

$$\|S_{\delta_{B_\delta}}\|_2^2 = p \cdot \text{trace}[\xi Q_q \xi] = p\Delta_h \Delta_v \cdot \text{trace}[Q_\delta]. \tag{4.7b}$$

Similarly, we get

$$\|S_{\delta_{C_\delta}}\|_2^2 = q \cdot \text{trace}[P_q] = q\Delta_h \Delta_v \cdot \text{trace}[P_\delta], \tag{4.7c}$$

and

$$\|S_{\delta_{D_\delta}}\|_2^2 = pq. \tag{4.7d}$$

Remark. In a similar manner, q-system counterpart to (4.7) may be obtained as follows:

$$\|S_{q_{A_q}}\|_1^2 \leq \text{trace}[P_q] \cdot \text{trace}[Q_q] = (\Delta_h \Delta_v)^2 \cdot \text{trace}[P] \cdot \text{trace}[\xi^{-1} Q \xi^{-1}]; \tag{4.8.a}$$

$$\|S_{q_{B_q}}\|_2^2 = p \cdot \text{trace}[Q_q] = p\Delta_h \Delta_v \cdot \text{trace}[\xi^{-1} Q \xi^{-1}]; \tag{4.8b}$$

$$\|S_{q_{C_q}}\|_2^2 = q \cdot \text{trace}[P_q] = q\Delta_h \Delta_v \cdot \text{trace}[P]; \tag{4.8.c}$$

$$\|S_{q_{D_q}}\|_2^2 = pq. \tag{4.8d}$$

Combining (4.5) with (4.7), we get the following upper bound for $M_{\delta_{\text{FXP}}}$:

$$\begin{aligned}
M_{\delta_{\text{FXP}}} \leq \overline{M}_{\delta_{\text{FXP}}} &\doteq (\text{trace}[P_q] + 1)(\text{trace}[\xi Q_q \xi] + 1) \\
&= (\Delta_h \Delta_v \cdot \text{trace}[P_\delta] + 1)(\Delta_h \Delta_v \cdot \text{trace}[Q_\delta] + 1).
\end{aligned} \tag{4.9}$$

Due to difficulties associated with minimization of $M_{\delta_{\text{FXP}}}$, it is customary to perform a minimization of $\overline{M}_{\delta_{\text{FXP}}}$. Hence, one attempts to characterize those realization $\{\tilde{A}_\delta, \tilde{B}_\delta, \tilde{C}_\delta, \tilde{D}_\delta\}$ that are 'bound optimal' with respect to $M_{\delta_{\text{FXP}}}$. For reasons of brevity, we do not attempt to perform this since it is exactly analogous to 2-D q-operator case (see [15], and references therein). For instance, one may show that any realization that is BL modulo an orthogonal nonsingular transformation is bound optimal with respect to $M_{\delta_{\text{FXP}}}$.

Remark. In a similar manner, q-system counterpart to (4.9) may be shown to be

$$\begin{aligned}
M_{q_{\text{FXP}}} \leq \overline{M}_{q_{\text{FXP}}} &\doteq (\text{trace}[P_q] + 1)(\text{trace}[Q_q] + 1) \\
&= (\Delta_h \Delta_v \cdot \text{trace}[P] + 1)(\Delta_h \Delta_v \cdot \text{trace}[\xi^{-1} Q \xi^{-1}] + 1).
\end{aligned} \tag{4.10}$$

However, it is instructive to note that, compared to a q-operator implementation, its δ-model implementation will always yield a smaller $\overline{M}_{\delta_{\text{FXP}}}$ whenever

$$\text{trace}[Q_q] > \text{trace}[\xi Q_q \xi] \;\Leftrightarrow\; (1-\Delta_h^2)\cdot\text{trace}[Q_q^{(1)}] + (1-\Delta_v^2)\cdot\text{trace}[Q_q^{(4)}] > 0. \tag{4.11}$$

Note that, with local reachability and observability of $\{A_\delta, B_\delta, C_\delta, D_\delta\}$ (and hence of $\{A_q, B_q, C_q, D_q\}$), positive definiteness of $Q_\delta^{(1)}$ and $Q_\delta^{(4)}$ (and hence of $Q_q^{(1)}$ and $Q_q^{(4)}$) are guaranteed. This implies strict positivity of $\text{trace}[Q_\delta^{(1)}]$ and $\text{trace}[Q_\delta^{(4)}]$ (and hence of $\text{trace}[Q_q^{(1)}]$ and $\text{trace}[Q_q^{(4)}]$). Thus, (4.11) is satisfied, that is, δ-operator implementation has superior coefficient sensitivity, whenever

$$\Delta_h < 1 \quad \text{and} \quad \Delta_v < 1. \tag{4.12}$$

FLP Arithmetic Case

For FLP arithmetic implementations, it is appropriate to define a *relative* sensitivity measure as

$$M_{\delta_{\text{FLP}}} = \|\tilde{S}_{\delta_{A_\delta}}\|_1^2 + \frac{1}{p}\|\tilde{S}_{\delta_{B_\delta}}\|_2^2 + \frac{1}{q}\|\tilde{S}_{\delta_{C_\delta}}\|_2^2 + \frac{1}{pq}\|\tilde{S}_{\delta_{D_\delta}}\|_2^2, \tag{4.13}$$

where

$$\begin{aligned} \tilde{S}_{\delta_{A_\delta}} &= \{\tilde{s}_{\delta_{A_\delta},ij}\} = a_{\delta_{ij}}\partial H_\delta/\partial a_{\delta_{ij}}; \qquad \tilde{S}_{\delta_{B_\delta}} = \{\tilde{s}_{\delta_{B_\delta},ij}\} = b_{\delta_{ij}}\partial H_\delta/\partial b_{\delta_{ij}}; \\ \tilde{S}_{\delta_{C_\delta}} &= \{\tilde{s}_{\delta_{C_\delta},ij}\} = c_{\delta_{ij}}\partial H_\delta/\partial c_{\delta_{ij}}; \qquad \tilde{S}_{\delta_{D_\delta}} = \{\tilde{s}_{\delta_{D_\delta},ij}\} = d_{\delta_{ij}}\partial H_\delta/\partial d_{\delta_{ij}}. \end{aligned} \tag{4.14}$$

Remark. Note that, in FLP, possible perturbation of a particular coefficient is approximately proportional to its nominal value. Hence, Li and Gevers [3], in addressing 1-D δ-system coefficient sensitivity, utilize a similar relative sensitivity measure.

LEMMA 4.3. The following bounds hold true:

$$\begin{aligned} \|\tilde{S}_{\delta_{A_\delta}}\|_p &\le \|A_\delta\|_F \cdot \|S_{\delta_{A_\delta}}\|_p; \qquad \|\tilde{S}_{\delta_{B_\delta}}\|_p =\le \|B_\delta\|_F \cdot \|S_{\delta_{B_\delta}}\|_p; \\ \|\tilde{S}_{\delta_{C_\delta}}\|_p &\le \|C_\delta\|_F \cdot \|S_{\delta_{C_\delta}}\|_p; \qquad \|\tilde{S}_{\delta_{D_\delta}}\|_p \le \|D_\delta\|_F \cdot \|S_{\delta_{D_\delta}}\|_p. \end{aligned}$$

Proof. Note that,

$$\begin{aligned} \|\tilde{S}_{\delta_{A_\delta}}\|_F^2 &= \sum\sum\|\tilde{s}_{\delta_{A_\delta},ij}\|_F^2 = \sum\sum\left\|a_{\delta_{ij}}\frac{\partial H_\delta}{\partial a_{\delta_{ij}}}\right\|_F^2 \\ &\le \sum\sum\|a_{\delta_{ij}}\|_F^2 \cdot \left\|\frac{\partial H_\delta}{\partial a_{\delta_{ij}}}\right\|_F^2 \le \sum\sum\|a_{\delta_{ij}}\|_F^2 \cdot \sum\sum\left\|\frac{\partial H_\delta}{\partial a_{\delta_{ij}}}\right\|_F^2 \\ &= \|A_\delta\|_F^2 \cdot \|S_{\delta_{A_\delta}}\|_F^2. \end{aligned}$$

Now, using Definition 4.1, one may verify the claim. ∎

Hence, substituting from (3.7), we get

$$M_{\delta_{\text{FLP}}} \le \|\xi^{-1}(A_q - I)\|_F^2 \cdot \|\xi S_{q_{A_q}}\|_1^2 + \frac{1}{p}\|\xi^{-1}B_q\|_F^2 \cdot \|\xi S_{q_{B_q}}\|_2^2 + \frac{1}{q}\|C_q\|_F^2 \cdot \|S_{q_{C_q}}\|_2^2 + \frac{1}{pq}\|D_q\|_F^2 \cdot \|S_{q_{D_q}}\|_2^2. \tag{4.15}$$

To proceed farther, let us assume $\Delta_h = \Delta_v = \Delta$ for convenience. Then, we get

$$\|\xi^{-1}(A_q - I)\|_F^2 = \frac{1}{\Delta^2}\|A_q - I\|_F^2; \qquad \|\xi^{-1}B_q\|_F^2 = \frac{1}{\Delta^2}\|B_q\|_F^2. \tag{4.16}$$

Combining (4.15) with (4.16), we get the following upper bound for $M_{\delta_{\text{FLP}}}$:

$$M_{\delta_{\text{FLP}}} \leq \overline{M}_{\delta_{\text{FLP}}} \doteq \|A_q - I\|_F^2 \cdot \text{trace}[P_q]\text{trace}[Q_q] + \|B_q\|_F^2 \cdot \text{trace}[Q_q] + \|C_q\|_F^2 \cdot \text{trace}[P_q] + \|D_q\|_F^2. \quad (4.17)$$

Again, we perform a minimization of $\overline{M}_{\text{FLP}}$.

Remark. In a similar manner, q-system counterpart to (4.17) may be shown to be

$$M_{q_{\text{FLP}}} \leq \overline{M}_{q_{\text{FLP}}} \doteq \|A_q\|_F^2 \cdot \text{trace}[P_q]\text{trace}[Q_q] + \|B_q\|_F^2 \cdot \text{trace}[Q_q] + \|C_q\|_F^2 \cdot \text{trace}[P_q] + \|D_q\|_F^2. \quad (4.18)$$

Hence, compared to a q-operator implementation, its δ-model implementation will yield a smaller $\overline{M}_{\delta_{\text{FLP}}}$ whenever

$$\|A_q - I_n\|_F^2 < \|A_q\|_F^2. \quad (4.19)$$

Clearly,

$$|\lambda_i[A_q] - 1| < |\lambda_i[A_q]|, \ \forall i = 1, \ldots, n \implies \|A_q - I_n\|_F^2 < \|A_q\|_F^2, \quad (4.20)$$

where $\lambda_i[A_q]$ denotes the i-th eigenvalue of A_q.

Remark. Li and Gevers [3] refer to the above region (where the eigenvalues of A_q should lie) as the *Middleton-Goodwin (MG) region.* They show that, for the 1-D case, δ-system offer superior performance (with respect to coefficient sensitivity) if the system eigenvalues lie within this MG region. It is well known that, high performance, high-Q, narrowband digital filters that operate under high sampling rates routinely satisfy this requirement. Hence, implementation of such filters via the proposed δ-model is expected to offer significant advantages over the conventional q-model.

V. Example

To illustrate the notions presented above, we consider a stable 5h-5v 2-D separable digital filter.

Computations

All numerical values are *displayed* via FORMAT SHORT E of MATLAB [22] which was used for all computations. Related references, equations, and MATLAB routines (displayed in `typewriter` font) associated with each result are indicated within ⟨angle brackets⟩. Note that, since system being considered has $A_q^{(3)} = \mathbf{0}$ (instead of $A_q^{(2)} = \mathbf{0}$), relevant equations must be appropriately modified.

Given q-model $\{A_q, B_q, C_q, D_q\}$. ⟨(2.1)⟩

$$A_q^{(1)} = \begin{bmatrix} 9.7288e-01 & 2.2120e-01 & -1.8087e-01 & 4.5533e-01 & -6.6164e-01 \\ -4.9620e-02 & 9.0641e-01 & 5.0270e-01 & -4.8744e-01 & 9.9640e-01 \\ -4.4045e-03 & -5.4572e-02 & 8.3446e-01 & 7.4341e-01 & -7.0769e-01 \\ -8.0077e-04 & -3.8212e-03 & -5.3688e-02 & 7.8877e-01 & 8.3955e-01 \\ -9.1701e-05 & -6.1563e-04 & -4.0265e-03 & -6.6197e-02 & 7.4779e-01 \end{bmatrix};$$

$$A_q^{(2)} = \begin{bmatrix} 6.5685e-04 & -3.5577e-04 & 6.5445e-05 & -4.7450e-06 & 7.6323e-08 \\ 3.7312e-04 & -2.0156e-04 & 3.7191e-05 & -2.8504e-06 & 9.4968e-08 \\ 8.3423e-05 & -4.5142e-05 & 8.3131e-06 & -6.1509e-07 & 1.3848e-08 \\ 1.0923e-05 & -6.0377e-06 & 1.0848e-06 & -4.3584e-08 & -1.0501e-08 \\ 1.3843e-06 & -8.0173e-07 & 1.3642e-07 & 5.0958e-09 & -4.8691e-09 \end{bmatrix};$$

$$A_q^{(3)} = [\mathbf{0}];$$

$$A_q^{(4)} = \begin{bmatrix} 9.7561e-01 & 5.1035e-02 & -4.7944e-03 & 9.6702e-04 & -1.0621e-04 \\ -2.0308e-01 & 9.1467e-01 & 5.7625e-02 & -4.1281e-03 & 6.5748e-04 \\ -1.5349e-01 & -4.6361e-01 & 8.4368e-01 & 5.4495e-02 & -3.9295e-03 \\ -4.0166e-01 & -4.3093e-01 & -7.0705e-01 & 7.9757e-01 & 6.1499e-02 \\ -5.9732e-01 & -9.2922e-01 & -6.9051e-01 & -8.3231e-01 & 7.5989e-01 \end{bmatrix};$$

$$B_q^{(1)} = [4.0636e-05 \quad 2.3637e-05 \quad 5.2062e-06 \quad 5.4906e-07 \quad 3.1533e-08]^T;$$

$$B_q^{(2)} = [8.2566e-01 \quad 2.0553e+00 \quad 4.2147e+00 \quad 7.7342e+00 \quad 1.3116e+01]^T;$$

$$C_q^{(1)} = [8.8401e-01 \quad -2.2366e+00 \quad 4.6091e+00 \quad -8.4165e+00 \quad 1.3754e+01];$$

$$C_q^{(2)} = [4.2897e-05 \quad -2.3778e-05 \quad 4.2589e-06 \quad -1.5122e-07 \quad -4.7764e-08];$$

$$D_q = [2.0931e-06].$$

Gramians P_q *and* Q_q. Note that, $P_q^{(2)} = Q_q^{(2)} = P_q^{(3)} = Q_q^{(3)} = \mathbf{0}$. Computation of $P_q^{(1)}$, $Q_q^{(1)}$, $P_q^{(4)}$, $Q_q^{(4)}$ is easily done using four Lyapunov equations. ⟨[12], `dlyap`⟩

$$P_q^{(1)} = \begin{bmatrix} 2.3413e-04 & 1.6529e-10 & 7.2257e-11 & -1.9205e-12 & -8.7578e-13 \\ 1.6529e-10 & 2.4873e-05 & -3.3583e-11 & 9.4761e-12 & -9.2682e-13 \\ 7.2257e-11 & -3.3583e-11 & 8.8928e-07 & -7.8387e-12 & -8.2826e-13 \\ -1.9205e-12 & 9.4761e-12 & -7.8387e-12 & 1.7156e-08 & -2.4931e-12 \\ -8.7578e-13 & -9.2682e-13 & -8.2826e-13 & -2.4931e-12 & 3.4901e-10 \end{bmatrix};$$

$$P_q^{(4)} = \begin{bmatrix} 1.5911e+01 & -8.0022e-04 & 3.9261e-04 & -2.8748e-04 & 1.1908e-03 \\ -8.0022e-04 & 3.1738e+01 & -1.8067e-03 & 2.0229e-03 & 5.8131e-04 \\ 3.9261e-04 & -1.8067e-03 & 8.9993e+01 & -1.2184e-02 & -4.3646e-03 \\ -2.8748e-04 & 2.0229e-03 & -1.2184e-02 & 3.2282e+02 & -3.9363e-01 \\ 1.1908e-03 & 5.8131e-04 & -4.3646e-03 & -3.9363e-01 & 1.1173e+03 \end{bmatrix};$$

$$Q_q^{(1)} = \begin{bmatrix} 1.6225e+01 & -1.1684e-04 & 5.0102e-05 & 2.7371e-04 & 7.3819e-04 \\ -1.1684e-04 & 3.4255e+01 & 1.4225e-04 & 1.4539e-05 & 2.2515e-05 \\ 5.0102e-05 & 1.4225e-04 & 1.0392e+02 & 1.3240e-02 & -1.9289e-02 \\ 2.7371e-04 & 1.4539e-05 & 1.3240e-02 & 3.8436e+02 & 7.0539e-01 \\ 7.3819e-04 & 2.2515e-05 & -1.9289e-02 & 7.0539e-01 & 1.2571e+03 \end{bmatrix};$$

$$Q_q^{(4)} = \begin{bmatrix} 2.3806e-04 & 6.5596e-06 & -1.5778e-06 & 1.9727e-07 & -1.3032e-08 \\ 6.5596e-06 & 2.7018e-05 & 1.0571e-06 & -1.3366e-07 & 1.0258e-08 \\ -1.5778e-06 & 1.0571e-06 & 1.0828e-06 & 3.9683e-08 & -3.5788e-09 \\ 1.9727e-07 & -1.3366e-07 & 3.9683e-08 & 2.2656e-08 & 7.4571e-10 \\ -1.3032e-08 & 1.0258e-08 & -3.5788e-09 & 7.4571e-10 & 4.2699e-10 \end{bmatrix}.$$

Hankel singular values Σ_q. Use Laub's algorithm to simultaneously diagonalize the pairs $\{P_q^{(1)}, Q_q^{(1)}\}$ and $\{P_q^{(4)}, Q_q^{(4)}\}$. ⟨[17], `chol`, `svd`, `dbalreal`⟩

$$\Sigma_q^{(1)} = \text{diag}\{\, 6.1633e-02 \quad 2.9190e-02 \quad 9.6133e-03 \quad 2.5679e-03 \quad 6.6237e-04 \,\};$$
$$\Sigma_q^{(4)} = \text{diag}\{\, 6.1613e-02 \quad 2.9240e-02 \quad 9.6278e-03 \quad 2.5201e-03 \quad 6.2671e-04 \,\}.$$

BL q-model $\{A_{qB}, B_{qB}, C_{qB}, D_{qB}\}$. ⟨`dbalreal`⟩

$$A_{qB}^{(1)} = \begin{bmatrix} 9.7288e-01 & -1.0477e-01 & 2.8226e-02 & 1.9100e-02 & 7.8007e-03 \\ 1.0477e-01 & 9.0641e-01 & 1.6563e-01 & 4.3176e-02 & 2.4797e-02 \\ 2.8226e-02 & -1.6563e-01 & 8.3445e-01 & -1.9981e-01 & -5.3485e-02 \\ -1.9098e-02 & 4.3169e-02 & 1.9980e-01 & 7.8859e-01 & -2.3575e-01 \\ 7.7946e-03 & -2.4778e-02 & -5.3426e-02 & 2.3570e-01 & 7.4798e-01 \end{bmatrix};$$

$$A_{qB}^{(2)} = \begin{bmatrix} 1.6387e-01 & -1.9665e-01 & 1.3359e-01 & -6.4934e-02 & 2.5945e-02 \\ -1.9656e-01 & 2.3525e-01 & -1.6009e-01 & 7.9605e-02 & -3.3830e-02 \\ -1.3338e-01 & 1.5991e-01 & -1.0869e-01 & 5.3270e-02 & -2.1779e-02 \\ 6.4945e-02 & -7.9570e-02 & 5.3310e-02 & -2.1310e-02 & 3.3012e-03 \\ -2.9278e-02 & 3.7620e-02 & -2.4427e-02 & 4.8708e-03 & 5.9844e-03 \end{bmatrix};$$

$$A_{qB}^{(3)} = [\,0\,];$$

$$A_{qB}^{(4)} = \begin{bmatrix} 9.7288e-01 & 1.0493e-01 & -2.8224e-02 & 1.9068e-02 & -6.9404e-03 \\ -1.0493e-01 & 9.0655e-01 & 1.6553e-01 & -4.3176e-02 & 2.2040e-02 \\ -2.8225e-02 & -1.6553e-01 & 8.3466e-01 & 1.9846e-01 & -4.7565e-02 \\ -1.9070e-02 & -4.3183e-02 & -1.9847e-01 & 7.8716e-01 & 2.1543e-01 \\ -6.9480e-03 & -2.2064e-02 & -4.7636e-02 & -2.1550e-01 & 7.9017e-01 \end{bmatrix};$$

$$B_{qB}^{(1)} = [\,6.5931e-04 \quad -8.0973e-04 \quad -5.4131e-04 \quad 2.1245e-04 \quad -4.3537e-05\,]^T;$$

$$B_{qB}^{(2)} = [\,5.4458e-02 \quad 6.5289e-02 \quad 4.4379e-02 \quad 2.1755e-02 \quad 8.9064e-03\,];$$

$$C_{qB}^{(1)} = [\,5.4485e-02 \quad 6.5289e-02 \quad -4.4331e-02 \quad -2.1760e-02 \quad -9.9934e-03\,];$$

$$C_{qB}^{(2)} = [\,6.5904e-04 \quad -8.0975e-04 \quad 5.4156e-04 \quad -2.0980e-04 \quad 2.3502e-05\,];$$

$$D_{qB} = [\,2.0931e-06\,].$$

Corresponding δ-model $\{A_\delta, B_\delta, C_\delta, D_\delta\}$. In FXP, selection of 'sampling times' Δ_h and Δ_v must be carefully done because they directly influence the number of bits required for integral and fractional portions of each coefficient. Actually, the relationships developed in Section 3 can be invaluable in determining suitable Δ_h and Δ_v so that the range of coefficient values of $\{A_{\delta B}, B_{\delta B}, C_{\delta B}, D_{\delta B}\}$ are acceptable. Let us select

$$\Delta_h = 5.0000e-01; \qquad \Delta_v = 2.5000e-01.$$

Note that, these are exactly representable (see remarks after (4.6)) and resulting coefficients of BL δ-model are each within $[-1, 1]$.

Of course, in FLP, such a difficulty does not usually arise because of the large dynamic range available. Hence, although one may choose smaller Δ_h and/or Δ_v, we will continue to use the same values.

Accordingly, we get the following δ-model: $\langle$(3.5), (3.7)$\rangle$

$$A_\delta^{(1)} = \begin{bmatrix} -5.4240e-02 & 4.4240e-01 & -3.6174e-01 & 9.1066e-01 & -1.3233e+00 \\ -9.9240e-02 & -1.8718e-01 & 1.0054e+00 & -9.7488e-01 & 1.9928e+00 \\ -8.8090e-03 & -1.0914e-01 & -3.3108e-01 & 1.4868e+00 & -1.4154e+00 \\ -1.6015e-03 & -7.6424e-03 & -1.0738e-01 & -4.2246e-01 & 1.6791e+00 \\ -1.8340e-04 & -1.2313e-03 & -8.0530e-03 & -1.3239e-01 & -5.0442e-01 \end{bmatrix};$$

$$A_\delta^{(2)} = \begin{bmatrix} 1.3137e-03 & -7.1154e-04 & 1.3089e-04 & -9.4900e-06 & 1.5265e-07 \\ 7.4624e-04 & -4.0312e-04 & 7.4382e-05 & -5.7008e-06 & 1.8994e-07 \\ 1.6685e-04 & -9.0284e-05 & 1.6626e-05 & -1.2302e-06 & 2.7696e-08 \\ 2.1846e-05 & -1.2075e-05 & 2.1696e-06 & -8.7168e-08 & -2.1002e-08 \\ 2.7686e-06 & -1.6035e-06 & 2.7284e-07 & 1.0192e-08 & -9.7382e-09 \end{bmatrix};$$

$$A_\delta^{(3)} = [\mathbf{0}];$$

$$A_\delta^{(4)} = \begin{bmatrix} -9.7560e-02 & 2.0414e-01 & -1.9178e-02 & 3.8681e-03 & -4.2484e-04 \\ -8.1232e-01 & -3.4132e-01 & 2.3050e-01 & -1.6512e-02 & 2.6299e-03 \\ -6.1396e-01 & -1.8544e+00 & -6.2528e-01 & 2.1798e-01 & -1.5718e-02 \\ -1.6066e+00 & -1.7237e+00 & -2.8282e+00 & -8.0972e-01 & 2.4600e-01 \\ -2.3893e+00 & -3.7169e+00 & -2.7620e+00 & -3.3292e+00 & -9.6044e-01 \end{bmatrix};$$

$$B_\delta^{(1)} = [8.1272e-05 \quad 4.7274e-05 \quad 1.0412e-05 \quad 1.0981e-06 \quad 6.3066e-08]^T;$$

$$B_\delta^{(2)} = [3.3026e+00 \quad 8.2212e+00 \quad 1.6859e+01 \quad 3.0937e+01 \quad 5.2464e+01]^T;$$

$$C_\delta^{(1)} = [8.8401e-01 \quad -2.2366e+00 \quad 4.6091e+00 \quad -8.4165e+00 \quad 1.3754e+01];$$

$$C_\delta^{(2)} = [4.2897e-05 \quad -2.3778e-05 \quad 4.2589e-06 \quad -1.5122e-07 \quad -4.7764e-08];$$

$$D_\delta = [2.0931e-06].$$

BL δ-model $\{A_{\delta B}, B_{\delta B}, C_{\delta B}, D_{\delta B}\}$. $\langle$Section III$\rangle$

$$A_{\delta B}^{(1)} = \begin{bmatrix} -5.4240e-02 & -2.0953e-01 & 5.6452e-02 & 3.8199e-02 & 1.5601e-02 \\ 2.0953e-01 & -1.8718e-01 & 3.3126e-01 & 8.6351e-02 & 4.9595e-02 \\ 5.6452e-02 & -3.3126e-01 & -3.3110e-01 & -3.9962e-01 & -1.0697e-01 \\ -3.8196e-02 & 8.6339e-02 & 3.9960e-01 & -4.2283e-01 & -4.7150e-01 \\ 1.5589e-02 & -4.9556e-02 & -1.0685e-01 & 4.7139e-01 & -5.0404e-01 \end{bmatrix};$$

$$A_{\delta B}^{(2)} = \begin{bmatrix} 4.6349e-01 & -5.5622e-01 & 3.7784e-01 & -1.8366e-01 & 7.3384e-02 \\ -5.5595e-01 & 6.6538e-01 & -4.5280e-01 & 2.2516e-01 & -9.5685e-02 \\ -3.7724e-01 & 4.5228e-01 & -3.0743e-01 & 1.5067e-01 & -6.1600e-02 \\ 1.8369e-01 & -2.2506e-01 & 1.5078e-01 & -6.0274e-02 & 9.3373e-03 \\ -8.2812e-02 & 1.0641e-01 & -6.9089e-02 & 1.3777e-02 & 1.6926e-02 \end{bmatrix};$$

$$A_{\delta B}^{(3)} = [\mathbf{0}];$$

$$A_{\delta B}^{(4)} = \begin{bmatrix} -1.0847e-01 & 4.1972e-01 & -1.1290e-01 & 7.6273e-02 & -2.7762e-02 \\ -4.1972e-01 & -3.7381e-01 & 6.6214e-01 & -1.7271e-01 & 8.8160e-02 \\ -1.1290e-01 & -6.6214e-01 & -6.6135e-01 & 7.9384e-01 & -1.9026e-01 \\ -7.6279e-02 & -1.7273e-01 & -7.9389e-01 & -8.5135e-01 & 8.6171e-01 \\ -2.7792e-02 & -8.8254e-02 & -1.9054e-01 & -8.6201e-01 & -8.3933e-01 \end{bmatrix};$$

$$B_{\delta B}^{(1)} = [9.3241e-04 \quad -1.1451e-03 \quad -7.6552e-04 \quad 3.0045e-04 \quad -6.1570e-05]^T;$$

$$B_{\delta B}^{(2)} = [\,1.0892e-01 \quad 1.3058e-01 \quad 8.8758e-02 \quad 4.3511e-02 \quad 1.7813e-02\,]^T;$$
$$C_{\delta B}^{(1)} = [\,7.7053e-02 \quad 9.2333e-02 \quad -6.2693e-02 \quad -3.0773e-02 \quad -1.4133e-02\,];$$
$$C_{\delta B}^{(2)} = [\,1.3181e-03 \quad -1.6195e-03 \quad 1.0831e-03 \quad -4.1961e-04 \quad 4.7004e-05\,];$$
$$D_{\delta B} = [\,2.0931e-06\,].$$

Simulations

Normalized frequency response of $\{A_q, B_q, C_q, D_q\}$ is

$$H_q\left(e^{j\omega_1}, e^{j\omega_2}\right) = C_q(I_z - A_q)^{-1}B_q + D_q\Big|_{\substack{z_1=e^{j\omega_1}\\ z_2=e^{j\omega_2}}},$$

whereas normalized frequency response of $\{A_\delta, B_\delta, C_\delta, D_\delta\}$ is

$$H_\delta\left((e^{j\omega_1}-1)/\Delta_h, (e^{j\omega_2}-1)/\Delta_v\right) = C_\delta(I_c - A_\delta)^{-1}B_\delta + D_\delta\Big|_{\substack{c_1=(e^{j\omega_1}-1)/\Delta_h\\ c_2=(e^{j\omega_2}-1)/\Delta_v}}.$$

Frequency responses are evaluated on the following grid:

$$\mathcal{G}^2 \doteq \{(\omega_1,\omega_2) \in \Re^2 : \omega_i = n_i \times \frac{\pi}{N},\ n_i = [-N:1:N],\ i = 1,2\},$$

and we selected $N = 32$.

For comparison purposes, the following measure was also evaluated:

$$E_{\max} \doteq \begin{cases} \max_{\mathcal{G}^2}\left|H\left(e^{j\omega_1}, e^{j\omega_2}\right) - \hat{H}\left(e^{j\omega_1}, e^{j\omega_2}\right)\right|, & \text{for } q\text{-models;} \\ \max_{\mathcal{G}^2}\left|H\left((e^{j\omega_1}-1)/\Delta_h, (e^{j\omega_2}-1)/\Delta_v\right) - \hat{H}\left((e^{j\omega_1}-1)/\Delta_h, (e^{j\omega_2}-1)/\Delta_v\right)\right|, & \text{for } \delta\text{-models.} \end{cases}$$

Here, H denotes the 'ideal' frequency response where each coefficient is represented in 'infinite' precision. $\hat{H}$ denotes the 'actual' frequency response where each coefficient is represented in finite precision.

Fig. (1). Plot shows the ideal frequency response. Note that, in 'infinite' precision, all realizations give identical results.

Fig. (2). Here, each coefficient is represented in FXP and its fractional part is truncated at different lengths. Integral part is represented exactly. Plot shows $E_{\max}$ versus number of fractional bits.

Remarks.

1. Advantage gained by BL model over given model is 6-7 bits.
2. Advantage gained by δ-model over its corresponding q-model is only 1 bit.
3. A small trace$[Q_q]$ implies trace$[Q_q]$+1 $\approx$ trace$[\xi Q_q \xi]$+1. Hence, from (4.9-10), no dramatic improvement in δ-model can be expected. This explains the modest gains in item 2 for this particular example.

Fig. (3). Here, each coefficient is represented in FXP and its total (integral+ fractional) number of bits is truncated at different lengths. Plot shows $E_{\max}$ versus total number of bits.

Remark.

1. This comparison is more realistic than what is in Fig. (2).
2. Only the BL realizations are shown; given systems' (q and δ) dynamic range are too large.

3. Advantage gained by δ-model over its corresponding q-model is 1-2 bits.
4. This modest improvement is due to large Δ_h and Δ_v being used. More dramatic improvements require smaller Δ_h and Δ_v (see (4.11)). However, this makes δ-model's coefficients to occupy a higher dynamic range. To circumvent this difficulty, careful scaling of filter coefficients must be performed. This is a research topic we are currently tackling.

Fig. (4). Here, each coefficient is represented in FLP and its number of mantissa bits is truncated at differene lengths. Plot shows $E_{\max}$ versus number of mantissa bits.

Remarks.

1. No apparent advantage gained by using corresponding BL model.
2. However, δ-models (BL or not) provide consistently better results with advantages of 3-4 bits.
3. Note that, $\|A_q - I_{10}\|_F = 2.7904e+00 < 3.8561e+00 = \|A_q\|_F$, $\|A_{qB} - I_{10}\|_F = 1.0502e+00 < 2.8612e+00 = \|A_{qB}\|_F$, and $\|A_{\delta B2q} - I_{10}\|_F = 1.1805e+00 < 2.9115e+00 = \|A_{\delta B2q}\|_F$. This explains the significant improvements shown by δ-models. For the particular example being considered, these differences between the two sides are not very high; if they were, more dramatic would be the improvement shown by the corresponding δ-model.

VI. Conclusion and Final Remarks

In this paper, we have developed a δ-operator based counterpart to the more conventional q-operator based Roesser local s.s. model. The motivation for this work lies in the superior finite wordlength properties exhibited by 1-D δ-operator based DT systems.

Corresponding notions of gramians and BL realization are proposed. By revealing the relationship between BL realizations of corresponding δ- and q-models, we have addressed computation of gramians and BL realizations as well. For both FXP and FLP implementations, conditions under which proposed δ-operator formulated systems behave better—with respect to coefficient sensitivity—than its q-operator counterpart are derived.

In the FXP case, δ-model is better whenever $\Delta_h < 1$ and $\Delta_v < 1$. However, this choice must be carefully done since, in FXP, δ-models tend to occupy a larger dynamic range. The authors are currently investigating the possibility of incorporating scaling of coefficients so that low values of Δ_h and Δ_v may be used to expose and exploit the advantages of δ-systems.

In the FLP case, such a limitation regarding dynamic range does not usually arise, and δ-models are better whenever the system matrix eigenvalues lie within the MG-region. This condition is typically true for high Q, narrowband digital filters operating under high sampling rates. We believe that, under these conditions, the proposed δ-models can yield significantly superior performance. In FLP, for comparative performance (with respect to coefficient sensitivity), δ-models require a shorter mantissa length. The ensuing implications regarding low power consumption, low cost and weight, and high speed cannot be overemphasized.

This work only addresses coefficient sensitivity issues. The authors are currently completing work regarding quantization noise properties of the δ-model developed, where, as in 1-D case, improvements over the corresponding q-model are expected.

We must mention that certain difficulties regarding limit cycles are inherent in δ-systems when FXP arithmetic is used [23]. However, this problem is, for all practical purposes, nonexistent in FLP arithmetic. Hence, in our opinion, for FLP high performance applications, the δ-model developed provides an extremely attractive solution that avoids numerical ill-conditioning typically associated with high speed q-systems.

References

[1] G.C. Goodwin, R.H. Middleton, and H.V. Poor, "High-speed digital signal processing and control," *Proc. IEEE,* vol. 80, pp. 240-259, 1992.

[2] G. Li and M. Gevers, "Roundoff noise minimization using delta-operator realizations," *IEEE Trans. Sig. Proc.,* vol. 41, pp. 629-637, Feb. 1993.

[3] G. Li and M. Gevers, "Comparative study of finite wordlength effects in shift and delta operator parameterizations," *Proc. 1990 IEEE Conf. Decision and Cont. (CDC'90),* pp. 954-959, Honolulu, HI, Dec. 1990.

[4] K. Premaratne, R. Salvi, N.R. Habib, and J.P. Le Gall, "Delta-operator formulated discrete-time equivalents of continuous-time systems," *IEEE Trans. Auto. Cont.,* vol. 39, pp. 581-585, Mar. 1994.

[5] R.H. Middleton and G.C. Goodwin, *Digital Control and Estimation: A Unified Approach,* Englewood Cliffs, NJ: Prentice-Hall, 1990.

[6] R. Vijayan, H.V. Poor, J.B. Moore, and G.C. Goodwin, "A Levinson-type algorithm for modeling fast-sampled data," *IEEE Trans. Auto. Cont.,* vol. 36, pp. 314-321, Mar. 1991.

[7] G. Likourezos, "Prolog to 'High-speed digital signal processing and control'," *Proc. IEEE,* vol. 80, pp. 238-239, 1992.

[8] R.P. Roesser, "A discrete state model for linear image processing," *IEEE Trans. Auto. Cont.,* vol. AC-20, pp. 1-10, Feb. 1975.

[9] E.I. Jury, "Stability of multidimensional systems and other related problems," Chapter 3 in *Multidimensional Systems, Techniques, and Applications,* New York, NY: Marcel Dekkar, 1986.

[10] J.W. Brewer, "Kronecker products and matrix calculus in system theory," *IEEE Trans. Circ. Syst.,* vol. CAS-25, pp. 772-781, Sept. 1978.

[11] K. Premaratne and A.S. Boujarwah, "Stability determination of two-dimensional delta-operator formulated discrete-time systems," to appear in *Multidim. Syst. Sig. Proc.,* 1994.

[12] K. Premaratne, E.I. Jury, and M. Mansour, "An algorithm for model reduction of 2-D discrete-time systems," *IEEE Trans. Circ. Syst.,* vol. CAS-37, pp. 1116-1132, Sept. 1990.

[13] W.-S. Lu, E.B. Lee, and Q.T. Zhang, "Model reduction for two-dimensional systems," *Proc. 1986 IEEE Int. Symp. Circ. Syst. (ISCAS'86),* vol. 1, pp. 79-82, 1986.

[14] T. Lin, M. Kawamata, and T. Higuchi, "A unified study on the roundoff noise in 2-D state-space digital filters," *IEEE Trans. Circ. Syst.,* vol. CAS-33, pp. 724-730, July 1986.

[15] T. Lin, M. Kawamata, and T. Higuchi, "Minimization of sensitivity of 2-D systems and its relation to 2-D balanced realizations," *The Trans. IEICE,* vol. E70, pp. 938-944, Oct. 1987; also in *Proc. 1987 IEEE Int. Symp. Circ. Syst. (ISCAS'87),* vol. 2, pp. 710-713, Philadelphia, PA, May 1987.

[16] W.-S. Lu and A. Antoniou, "Synthesis of 2-D state-space fixed-point digital filter structures with minimum roundoff noise," *IEEE Trans. Circ. Syst.,* vol. CAS-33, pp. 965-973, Oct. 1986.

[17] A.J. Laub, M.T. Heath, C.C. Paige, and R.C. Ward, "Computation of system balancing transformations and other applications of simultaneous diagonalization algorithms," *IEEE Trans. Auto. Cont.,* vol. AC-32, pp. 115-122, Feb. 1987.

[18] W.-S. Lu, H.-P. Wang, and A. Antoniou, "An efficient method for the evaluation of the controllability and observability gramians of 2-D digital filters and systems," *IEEE Trans. Circ. Syst.—II. Anal. Dig. Sig. Proc.,* vol. 39, pp. 695-704, Oct. 1992.

[19] V. Tavsanoglu and L. Thiele, "Optimal design of state-space digital filters by simultaneous minimization of sensitivity and roundoff noise," *IEEE Trans. Circ. Syst.,* vol. CAS-31, pp. 884-888, Oct. 1984.

[20] W.J. Lutz and S.L. Hakimi, "Design of multi-input multi-output systems with minimum sensitivity," *IEEE Trans. Circ. Syst.,* vol. 35, pp. 1114-1122, Sept. 1988.

[21] G.H. Golub and C.F. Van Loan, *Matrix Computations,* Baltimore, MD: John Hopkins University Press, 1983.

[22] *MATLAB,* ver. 4.2a, Natick, MA: The MathWorks Inc.

[23] K. Premaratne and P.H. Bauer, "Limit cycles and asymptotic stability of delta-operator systems in fixed-point arithmetic," *Proc. 1994 IEEE Int. Symp. Circ. Syst. (ISCAS'94),* London, UK, vol. 2, pp. 461-464, May 1994.

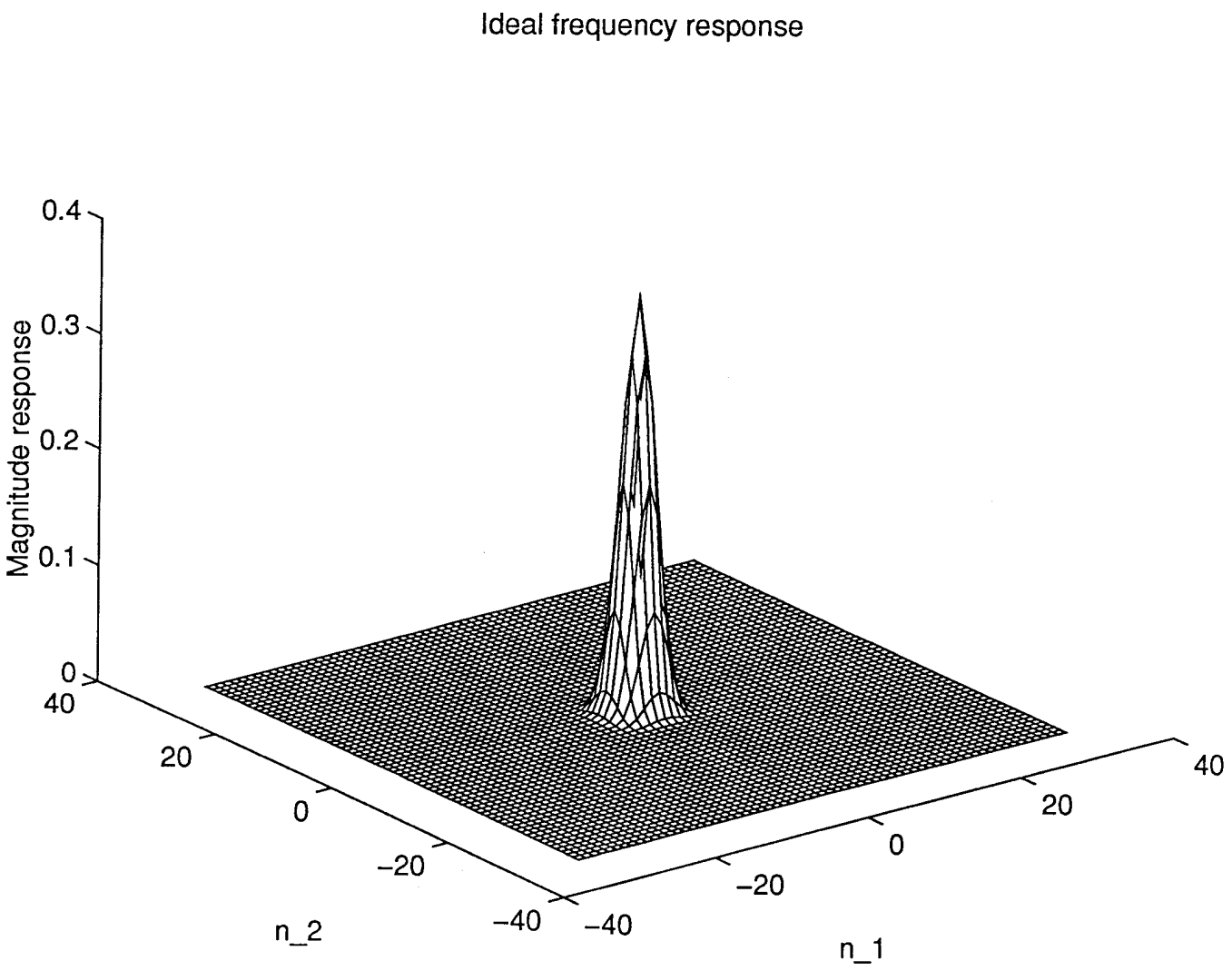

Ideal frequency response
Magnitude response
0.4
0.3
0.2
0.1
0
40
20
0
−20
−40
n_2
−40
−20
0
20
40
n_1

Figure (1)

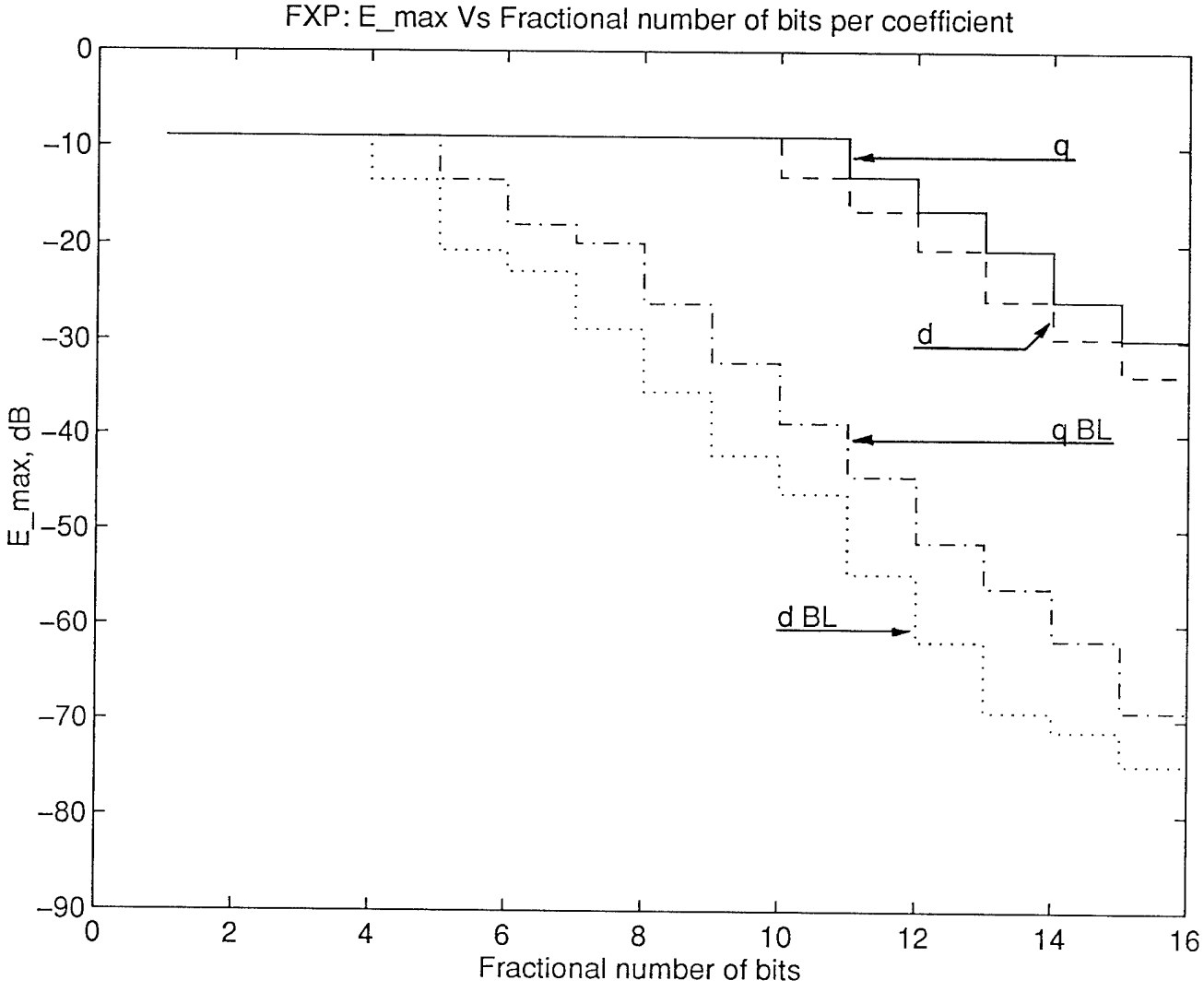

FXP: E_max Vs Fractional number of bits per coefficient
E_max, dB
Fractional number of bits
0
-10
-20
-30
-40
-50
-60
-70
-80
-90
0
2
4
6
8
10
12
14
16
q
d
q BL
d BL

Figure (2)

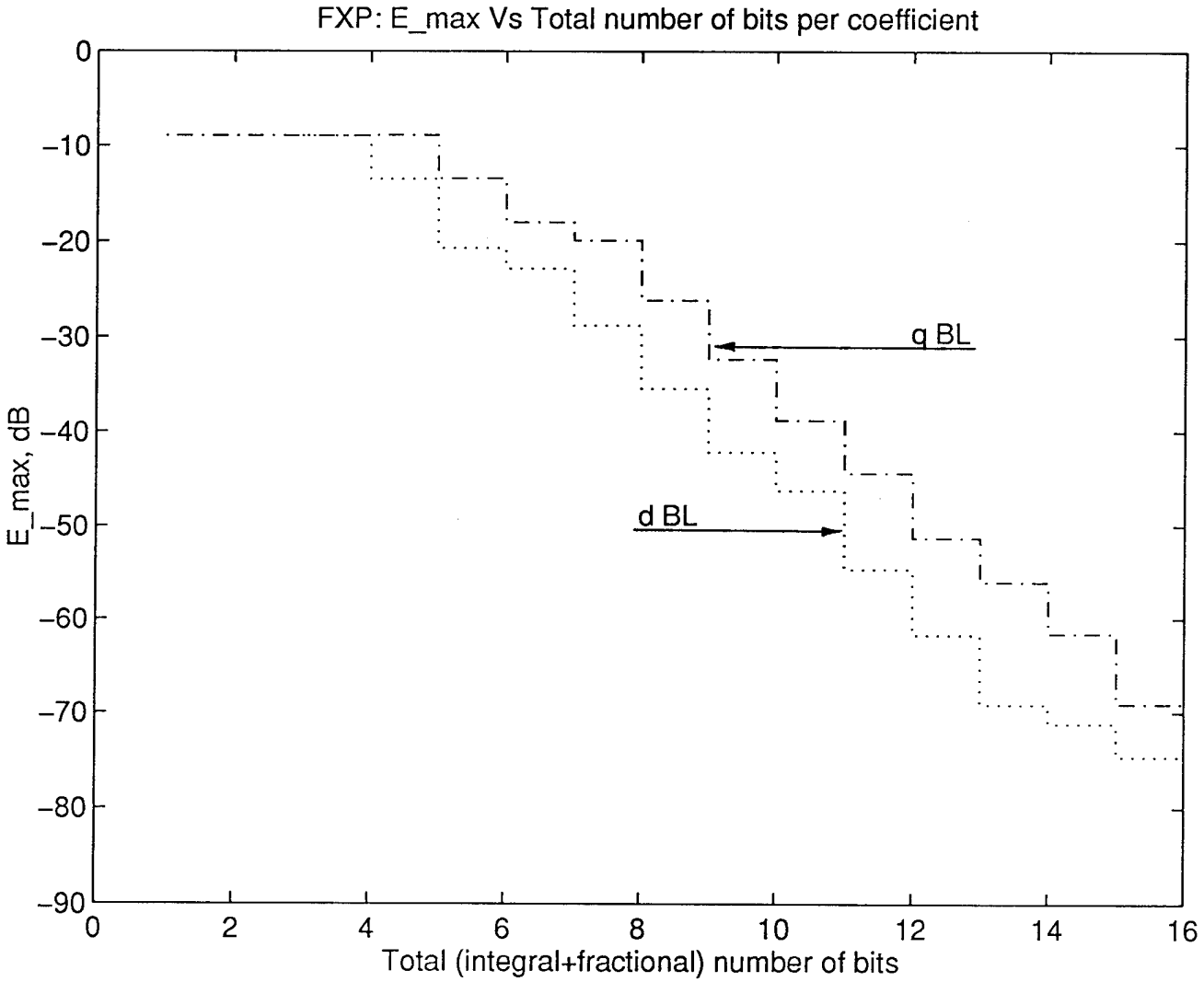

FXP: E_max Vs Total number of bits per coefficient
E_max, dB
0
−10
−20
−30
−40
−50
−60
−70
−80
−90
0
2
4
6
8
10
12
14
16
Total (integral+fractional) number of bits
q BL
d BL

Figure (3)

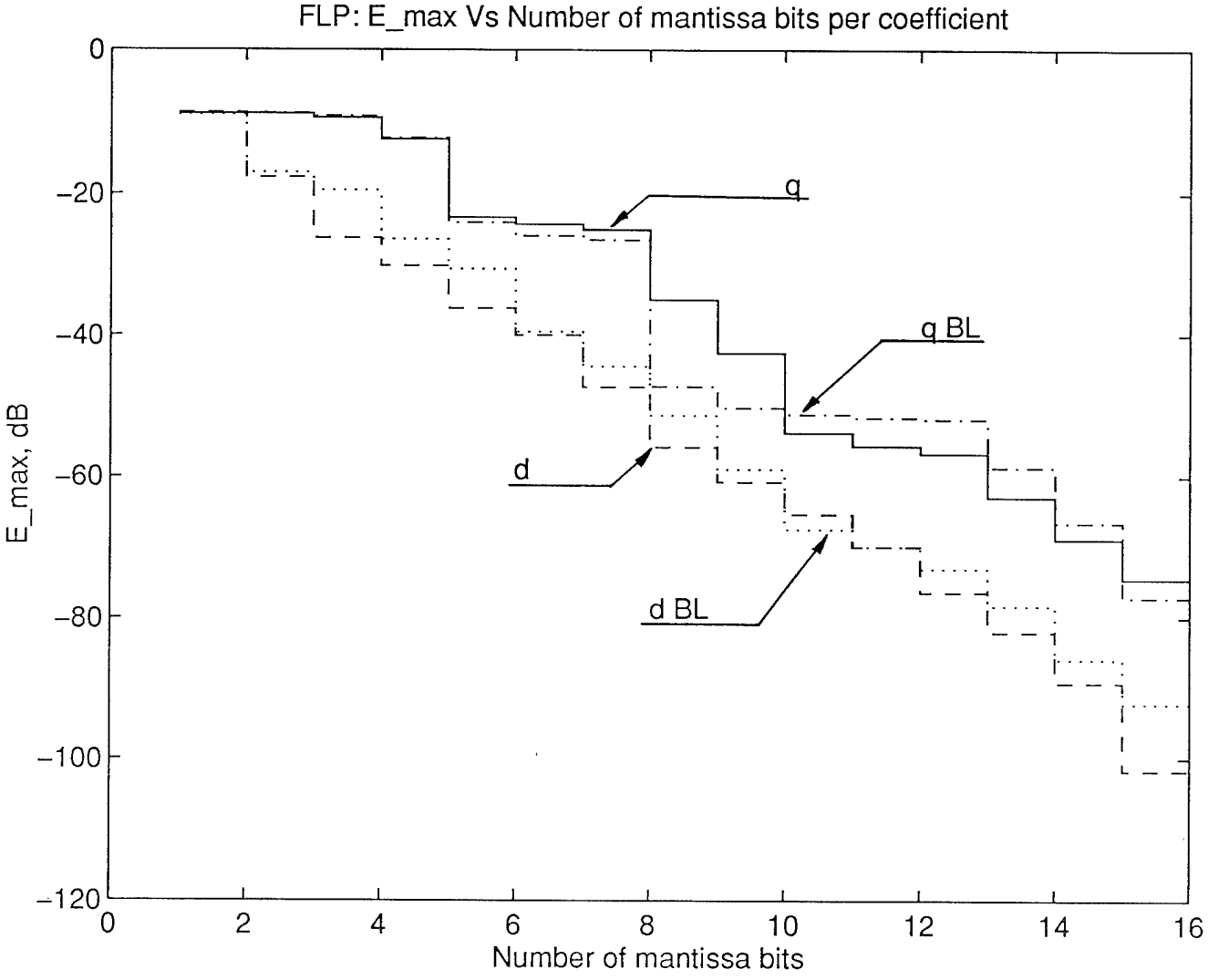

FLP: E_max Vs Number of mantissa bits per coefficient
E_max, dB
Number of mantissa bits
0
−20
−40
−60
−80
−100
−120
0
2
4
6
8
10
12
14
16
q
q BL
d
d BL

Figure (4)

Proceedings of the IEEE SOUTHEASTCON'94, Miami, FL, pp. 26-29, April 1994

FIXED-POINT IMPLEMENTATION OF MULTI-DIMENSIONAL DELTA-OPERATOR FORMULATED DISCRETE-TIME SYSTEMS: DIFFICULTIES IN CONVERGENCE

Peter H. Bauer, PhD
Department of Electrical Engineering
Laboratory of Image and Signal Analysis
University of Notre Dame
Notre Dame, IN 46556

Kamal Premaratne, PhD
Department of Electrical and
Computer Engineering
University of Miami
Coral Gables, FL 33124

Abstract— In this paper, the convergence properties of linearly stable multi-dimensional systems are investigated for the case of delta-operator implementations in fixed-point format. It is shown that zero-convergence is almost never achieved, if the sampling time is small. Using a one-dimensional analysis, it is demonstrated that zero-convergence cannot be guaranteed along the axis of the first hyper-quadrant for a first hyper-quadrant causal system. This limits the use of delta-operators for solving partial differential equations in discrete time with fixed-point arithmetic.

I. INTRODUCTION

Delta-operator (or, δ-operator) implementations of discrete-time systems have been the topic of a number of research papers within the last decade. A comprehensive treatment of the properties of δ-operator implementations can be found in [1]. It is well known that δ-operators outperform shift-operators (or, q-operators) in terms of their finite wordlength properties [2]. In particular, its quantization noise and sensitivity properties make the δ-operator an interesting alternative to the q-operator in areas such as digital control, digital signal processing, and generally discrete-time simulation of dynamical systems described by differential equations [1], [3].

In this paper, we will perform a deterministic analysis of the finite wordlength properties of multi-dimensional (m-D) δ-operator implemented discrete-time systems. In particular, we will investigate the zero-convergence of δ-operator fixed-point implementations of one-dimensional (1-D) and m-D systems. Although it is of vital importance, this problem has not been investigated thus far in the literature. After all, asymptotic stability and convergence to the true equilibrium points are some of the most fundamental requirements for any discrete-time system realization.

This article is organized in the following way: Section II introduces the notation. The m-D δ-operator model will be introduced and briefly discussed. This section will also provide the problem formulation. Section III provides necessary 1-D stability conditions for m-D first hyper-quadrant causal systems with nonlinearities. Using these necessary conditions, section IV provides a stability and convergence analysis for m-D systems. It will be shown that the resulting 1-D systems cannot ensure zero-convergence. Section V contains concluding remarks.

II. NOTATION AND PROBLEM FORMULATION

The m-D Roesser model has the following δ-operator formulation [4]:

$$\begin{bmatrix} \delta^{(1)}[\mathbf{x}^{(1)}](\mathbf{n}) \\ \vdots \\ \delta^{(m)}[\mathbf{x}^{(m)}](\mathbf{n}) \end{bmatrix} = \begin{bmatrix} A^{\delta}_{11} & \cdots & A^{\delta}_{1m} \\ \vdots & \ddots & \vdots \\ A^{\delta}_{m1} & \cdots & A^{\delta}_{mm} \end{bmatrix} \begin{bmatrix} \mathbf{x}^{(1)}(\mathbf{n}) \\ \vdots \\ \mathbf{x}^{(m)}(\mathbf{n}) \end{bmatrix} + \begin{bmatrix} B^{\delta}_{1} \\ \vdots \\ B^{\delta}_{m} \end{bmatrix} \mathbf{u}(\mathbf{n}); \quad (1)$$

$$\begin{bmatrix} q^{(1)}[\mathbf{x}^{(1)}](\mathbf{n}) \\ \vdots \\ q^{(m)}[\mathbf{x}^{(m)}](\mathbf{n}) \end{bmatrix} = \begin{bmatrix} \mathbf{x}^{(1)}(\mathbf{n}) \\ \vdots \\ \mathbf{x}^{(m)}(\mathbf{n}) \end{bmatrix} + \Delta \cdot \begin{bmatrix} \delta^{(1)}[\mathbf{x}^{(1)}](\mathbf{n}) \\ \vdots \\ \delta^{(m)}[\mathbf{x}^{(m)}](\mathbf{n}) \end{bmatrix}. \quad (2)$$

The input-state equations in (1) and (2) describe a first hyper-quadrant causal m-D system with a uniform sampling period of Δ in all directions. The operators $q^{(i)}$ and $\delta^{(i)}$ represent the shift- and delta-operator in the direction specified by the axis n_i. In particular

$$q^{(i)}[\mathbf{x}^{(i)}](\mathbf{n}) = \mathbf{x}^{(i)}(n_1, \ldots, n_{i-1}, n_i + 1, n_{i+1}, \ldots, n_m) \quad (3a)$$

$$\delta^{(i)}[\mathbf{x}^{(i)}](\mathbf{n})$$

$$= \frac{1}{\Delta}(\mathbf{x}^{(i)}(n_1,\ldots,n_{i-1},n_i+1,n_{i+1},\ldots,n_m) - \mathbf{x}^{(i)}(\mathbf{n})). \quad (3b)$$

Here, $(\mathbf{n}) \doteq (n_1,\ldots,n_m)$ denotes a point in the first hyper-quandrant, $\mathbf{x}^{(i)}(\mathbf{n})$ is the portion of the state vector propagating in the direction specified by the axis n_i, $u(\mathbf{n})$ is the m-D input vector, and A_{ij}^{δ} and B_i^{δ}, for$i = 1,\ldots,m$, $j = 1,\ldots,m$, are the submatrices of the system and input matrices, respectively.

If (1) is realized in fixed-point arithmetic, it takes the following form under zero-input conditions:

$$\begin{bmatrix} \delta^{(1)}[\mathbf{x}^{(1)}](\mathbf{n}) \\ \vdots \\ \delta^{(1)}[\mathbf{x}^{(1)}](\mathbf{n}) \end{bmatrix} = \mathbf{Q}\left\{ \begin{bmatrix} A_{11}^{\delta} & \cdots & A_{1m}^{\delta} \\ \vdots & \ddots & \vdots \\ A_{m1}^{\delta} & \cdots & A_{mm}^{\delta} \end{bmatrix} \begin{bmatrix} \mathbf{x}^{(1)}(\mathbf{n}) \\ \vdots \\ \mathbf{x}^{(m)}(\mathbf{n}) \end{bmatrix} \right\} \quad (4)$$

where $\mathbf{Q}\{\mathbf{x}\} = \begin{pmatrix} Q\{x_1\} \\ \vdots \\ Q\{x_m\} \end{pmatrix}$ with $\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix}$.

Equation (4) assumes quantization after summation; since practically all modern DSP machines implement this quantization scheme, we utilize this. The vector-valued quantization nonlinearity $\mathbf{Q}\{\cdot\}$ may represent any one of the conventional schemes, viz., magnitude truncation, magnitude rounding, two's complement truncation, and two's complement rounding.

Equation (2) can be implemented in two different forms:

$$\begin{bmatrix} q^{(1)}[\mathbf{x}^{(i)}](\mathbf{n}) \\ \vdots \\ q^{(m)}[\mathbf{x}^{(m)}](\mathbf{n}) \end{bmatrix} = \begin{bmatrix} \mathbf{x}^{(1)}(\mathbf{n}) \\ \vdots \\ \mathbf{x}^{(m)}(\mathbf{n}) \end{bmatrix} + \mathbf{Q}\left\{ \Delta \cdot \begin{bmatrix} \delta^{(1)}[\mathbf{x}^{(1)}](\mathbf{n}) \\ \vdots \\ \delta^{(m)}[\mathbf{x}^{(m)}](\mathbf{n}) \end{bmatrix} \right\} \quad (5)$$

or

$$\begin{bmatrix} q^{(1)}[\mathbf{x}^{(1)}](\mathbf{n}) \\ \vdots \\ q^{(m)}[\mathbf{x}^{(m)}](\mathbf{n}) \end{bmatrix} = \mathbf{Q}\left\{ \begin{bmatrix} \mathbf{x}^{(1)}(\mathbf{n}) \\ \vdots \\ \mathbf{x}^{(m)}(\mathbf{n}) \end{bmatrix} + \Delta \cdot \begin{bmatrix} \delta^{(1)}[\mathbf{x}^{(1)}](\mathbf{n}) \\ \vdots \\ \delta^{(m)}[\mathbf{x}^{(m)}](\mathbf{n}) \end{bmatrix} \right\} \quad (6)$$

Equation (5) corresponds to quantization after multiplication, whereas (6) corresponds to quantization after addition. In contrast to (1), for (2), it is not obvious which of the two forms stated above is preferable.

The following definition for asymptotic stability [5] will be used throughout this paper.

Definition. An m-D first hyper-quadrant causal discrete-time system is asymptotically stable under all finitely extended bounded input signals $u(\mathbf{n})$ where

$$|u(\mathbf{n})| \leq S, \quad \text{for} \quad n_1 + \cdots + n_m \leq D; \quad (7)$$

$$u(\mathbf{n}) = 0, \quad \text{for} \quad n_1 + \cdots + n_m > D, \quad (8)$$

if all the states of the m-D discrete-time system asymptotically reach zero for $n_1 + \cdots + n_m \to \infty$. Here, $n_\nu \geq 0$, $\nu = 1,\ldots,m$, S is a nonnegative real number, and D is a positive integer.

Since the fixed-point systems considered are in fact finite state machines, the condition

$$\begin{pmatrix} \mathbf{x}^{(1)}(\mathbf{n}) \\ \vdots \\ \mathbf{x}^{(m)}(\mathbf{n}) \end{pmatrix} \to \mathbf{0},$$

for $n_1 + \cdots + n_m \to \infty$, $n_\nu \geq 0$, $\nu = 1,\ldots,m$, can be strengthened to

$$\begin{pmatrix} \mathbf{x}^{(1)}(\mathbf{n}) \\ \vdots \\ \mathbf{x}^{(m)}(\mathbf{n}) \end{pmatrix} = \mathbf{0},$$

for all points $n_1 + \cdots + n_m \geq c$, $n_\nu \geq 0$, $\nu = 1,\ldots,m$, where c is some finite integer.

Problem Formulation. Analyze the asymptotic zero-convergence of the state response of systems in (4,5) and (4,6) under the assumption that the underlying linear system is asymptotically stable.

III. NECESSARY CONDITIONS FOR GLOBAL ASYMPTOTIC STABILITY OF m-D SYSTEMS

In this section, we present some necessary conditions for stability of a first hyper-quadrant causal m-D discrete-time system represented in its Roesser local state-space model in (1,2). These necessary conditions are formulated in terms of 1-D conditions. This theorem follows directly from a result in [6] which was formulated for q-operator implemented discrete-time systems. The proof of the theorem rests on the fact that a first hyper-quadrant m-D system can be described by a 1-D system for those locations that are along the m coordinate axes of the boundary of the hyper-quadrant. Reformulating the result in [6] for δ-operator systems produces the following theorem:

Theorem 1.

(a) A necessary condition for global asymptotic stability of the system in (4,5) is that each of the following 1-D systems in (9,10) is globally asymptotically stable:

$$\delta^{(i)}[\mathbf{x}^{(i)}](n_i) = \mathbf{Q}\left\{[A_{ii}^{\delta}]\mathbf{x}^{(i)}(n_i)\right\}; \tag{9}$$

$$q^{(i)}[\mathbf{x}^{(i)}](n_i) = \mathbf{x}^{(i)}(n_i) + \mathbf{Q}\left\{\Delta \cdot \delta^{(i)}[\mathbf{x}^{(i)}](n_i)\right\} \tag{10}$$

where $i = 1, \ldots, m$.

(b) A necessary condition for global asymptotic stability of the system in (4,6) is that each of the following in 1-D systems in (11,12) is globally asymptotically stable:

$$\delta^{(i)}[\mathbf{x}^{(i)}](n_i) = \mathbf{Q}\left\{[A_{ii}^{\delta}]\mathbf{x}^{(i)}(n_i)\right\}; \tag{11}$$

$$q^{(i)}[\mathbf{x}^{(i)}](n_i) = \mathbf{Q}\left\{\mathbf{x}^{(i)}(n_i) + \Delta \cdot \delta^{(i)}[\mathbf{x}^{(i)}](n_i)\right\} \tag{12}$$

where $i = 1, \ldots, m$.

Proof. For a detailed proof, and generalizations to higher sub-dimensional systems, the reader is referred to [6]. ■

Theorem 1 can be viewed as an extension of the concept of practical BIBO stability to asymptotic stability of nonlinear systems. It is particularly useful in proving instability in m-D nonlinear systems.

IV. NECESSARY CONDITIONS FOR GLOBAL ASYMPTOTIC STABILITY OF 1-D SYSTEMS

Let us rewrite (9), (10), and (12) as 1-D matrix equations of order K. In this case, (9), (10), and (12) yield (13), (14), and (15), respectively:

$$\begin{bmatrix} \delta[x_1](n) \\ \vdots \\ \delta[x_K](n) \end{bmatrix} = \mathbf{Q}\left\{ \begin{bmatrix} a_{11}^{\delta} & \cdots & a_{1K}^{\delta} \\ \vdots & \ddots & \vdots \\ a_{n1}^{\delta} & \cdots & a_{KK}^{\delta} \end{bmatrix} \begin{bmatrix} x_1(n) \\ \vdots \\ x_K(n) \end{bmatrix} \right\}; \tag{13}$$

$$\begin{bmatrix} x_1(n+1) \\ \vdots \\ x_K(n+1) \end{bmatrix} = \begin{bmatrix} x_1(n) \\ \vdots \\ x_K(n) \end{bmatrix} + \mathbf{Q}\left\{ \Delta \cdot \begin{bmatrix} \delta[x_1](n) \\ \vdots \\ \delta[x_K](n) \end{bmatrix} \right\}; \tag{14}$$

$$\begin{bmatrix} x_1(n+1) \\ \vdots \\ x_K(n+1) \end{bmatrix} = \mathbf{Q}\left\{ \begin{bmatrix} x_1(n) \\ \vdots \\ x_K(n) \end{bmatrix} + \Delta \cdot \begin{bmatrix} \delta[x_1](n) \\ \vdots \\ \delta[x_K](n) \end{bmatrix} \right\}. \tag{15}$$

Now, we are in a position to formulate the second theorem which presents a necessary condition for stability of 1-D systems.

Theorem 2. A necessary condition for global asymptotic stability of the system in (13,14) or (13,15) is given by

$$\Delta \geq 0.5, \quad \text{for magnitude rounding;}$$
$$\Delta \geq 1, \quad \text{for truncating.}$$

Proof. For global asymptotic stability of (13,14), it is necessary that

$$\mathbf{Q}\left\{ \Delta \cdot \begin{bmatrix} \delta[x_1](n) \\ \vdots \\ \delta[x_K](n) \end{bmatrix} \right\} \neq \mathbf{0}, \tag{16}$$

for any $\begin{pmatrix} x_1(n) \\ \vdots \\ x_K(n) \end{pmatrix} \neq \mathbf{0}$.

First, we will address the case of magnitude rounding. Obviously, condition (16) is violated if, for $x_\nu \neq 0$,

$$|\Delta \cdot \delta[x_\nu](n)| < \frac{\ell}{2}, \quad \text{for} \quad \nu = 1, \ldots, K, \tag{17}$$

where ℓ is the quantization step. Expressing the sampling time Δ as an integer multiple of ℓ, we have

$$\Delta = I \cdot \ell, \tag{18}$$

where I is some (typically small) positive integer. With (17) and (18), we obtain the following condition for instability:

$$|\delta[x_v](n)| < \frac{1}{2I}, \ \nu = 1, \ldots, m, \tag{19}$$

for $x_\nu \neq 0$, $\nu = 1, \ldots, m$.

Condition (19) is not satisfied for any nonzero value of x_ν (that is, the condition for instability is not satisfied) if $\ell \geq 1/2I$, or equivalently,

$$\Delta \geq \frac{1}{2}. \tag{20}$$

This proves the theorem for magnitude rounding.

For the case of magnitude truncating, (17) takes the form

$$|\Delta \cdot \delta[x_v](n)| < \ell, \quad \text{for} \quad \nu = 1, \ldots, K. \tag{21}$$

Therefore, (19) becomes

$$|\delta[x_v](n)| < \frac{1}{I}. \tag{22}$$

This finally yields

$$\Delta \geq 1. \tag{23}$$

For two's complement, (17) takes the form

$$0 \leq \Delta \cdot \delta[x_\nu](n) < \ell, \quad \text{for} \quad \nu = 1, \ldots, K. \tag{24}$$

This results in

$$0 \leq \delta[x_v](n) < \frac{1}{I}, \tag{25}$$

and consequently, $\Delta \geq 1$. This proves the theorem for the system in (13,14). A similar argument can be used for the system in (13,15) by considering the cases for which

$$\mathbf{Q}\left\{\begin{bmatrix} x_1(n) \\ \vdots \\ x_K(n) \end{bmatrix} + \Delta \cdot \begin{bmatrix} \delta[x_1](n) \\ \vdots \\ \delta[x_K](n) \end{bmatrix}\right\} = \mathbf{Q}\left\{\begin{bmatrix} x_1(n) \\ \vdots \\ x_K(n) \end{bmatrix}\right\}, \tag{26}$$

for nonzero state vectors. ■

We can now combine Theorems 1 and 2 to formulate a necessary condition for stability of m-D first hyperquadrant causal δ-operator formulations of the generalized Roesser model.

Corollary 3. A necessary condition for global asymptotic stability of the m-D systems in (4,5) or (4,6) is

$$\Delta \geq 0.5, \quad \text{for magnitude rounding;}$$
$$\Delta \geq 1, \quad \text{for truncatiing.}$$

Proof. The proof follows from Theorems 1 and 2. ■

Comments.

1. Theorem 2 and Corollary 3 are also essentially applicable to the case where the sampling time varies with the direction of propagation. In this case, the inequalities in Theorem 2 and Corollary 3 would have to be replaced by

   $$\Delta_i \geq 0.5, \quad \text{for magnitude rounding;}$$
   $$\Delta_i \geq 1, \quad \text{for truncating,}$$

   for $i = 1, \ldots, m$.
2. Most of the previous results on the superior finite wordlength properties of δ-operators depend on choosing a very small sampling time Δ. In such a case, Theorem 2 and Corollary 3 show that the system response will not converge to zero for the unforced case.
3. Our analysis is limited to the zero-input case for which DC limit cycles were used to derive conditions for non-convergence. If one includes other types of limit cycles in the analysis, the requirements for Δ may become even more severe.
4. Theorem 2 and Corollary 3 show that fixed-point implementations of 1-D and m-D δ-operator systems *cannot be realized limit cycle free, if good coefficient sensitivity and quantization noise measures have to be achieved.* See also [7].

V. CONCLUSION

In this paper, it was shown that fixed-point implementations of 1-D and m-D δ-operator systems are not limit cycle free even if the underlying linear system is stable and the sampling time is chosen small. This nonconvergent behavior can be explained by the quantization of the δ-term to zero which leaves the state vector unchanged. The smaller the sampling time, the more severe this effect is. Therefore, the practical value of δ-operators for fixed-point implementations of 1-D and m-D systems is questionable. There are however indications that this effect is much less severe in floating-point implementations.

δ-operator implemented discrete-time systems represent a class of systems where the quantization noise at the output can be small compared to other realizations. However, as was shown above, such realizations will invariably exhibit limit cycle, that is, highly correlated quantization noise, behavior. Therefore, in this case, typical measures for quantization noise are of very limited use for obtaining any insight into the likelihood of limit cycles and vice versa.

ACKNOWLEDGEMENT

This work was partially supported by a grant from the Office of Naval Research (ONR).

REFERENCES

[1] G.C. Goodwin, R.H. Middleton, and H.V. Poor, "High-speed digital signal processing and control," *Proceedings of the IEEE,* vol. 80, no. 2, pp. 240-259, Feb. 1992.

[2] R.H. Middleton and G.C. Goodwin, "Improved finite wordlength characteristics in digital control using delta operators," *IEEE Transactions on Automatic Control,* vol. 31, pp. 1015-1021, Nov. 1986.

[3] G.Li and M. Gevers, "Comparative study of finite wordlength effects in shift and delta operator parameterization," *Proceedings of the IEEE Conference on Decision and Control (CDC'90),* vol. 2, pp. 954-959, Honolulu, HI, 1990.

[4] K. Premaratne, J. Suarez, M.M. Ekanayake, and P.H. Bauer, "Delta-operator formulated implementation of two-dimensional discrete-time systems," in preparation.

[5] P. Bauer, "Finite wordlength effects in m-D digital filters with singularities on the stability boundary," *IEEE Transactions on Signal Processing,* vol. 40, no. 4, pp. 894-900, Apr. 1992.

[6] P. Bauer, "A set of necessary stability conditions for m-D nonlinear digital filters," to appear in *Circuits, Systems and Signal Processing,* 1994.

[7] K. Premaratne and P.H. Bauer, "Limit cycles and asymptotic stability of delta-operator systems in fixed-point arithmetic," submitted to be presented at the *1994 IEEE Symposium on Circuits and Systems (ISCAS'94),* London, UK, 1994.

Proceedings of the 1994 IEEE International Symposium on Circuits and Systems (ISCAS'94), London, UK, pp. 461-464, May-June 1994.

Limit cycles and asymptotic stability of delta-operator formulated discrete-time systems implemented in fixed-point arithmetic

Kamal Premaratne
Department of Electrical and Computer Engineering
University of Miami
Coral Gables, FL 33124
USA
(+1) 305-284-4051
kprema@umiami.ir.miami.edu

Peter H. Bauer
Department of Electrical Engineering
Laboratory of Image and Signal Analysis
University of Notre Dame
Notre Dame, IN 46556
USA
(+1) 219-631-8015
pbauer@mars.ee.nd.edu

ABSTRACT

This paper analyzes the problem of global asymptotic stability of delta-operator formulated discrete-time systems implemented in fixed-point arithmetic. It is shown that the free response of such a system tends to produce period one limit cycles if conventional quantization arithmetic schemes are used. Explicit necessary conditions for global asymptotic stability are derived, and these demonstrate that, in almost all cases, fixed-point arithmetic does not allow for global asymptotic stability in delta-operator formulated discrete-time systems that use a short sampling time.

I. INTRODUCTION

Recently, discrete-time systems formulated with the incremental difference operator (or, δ-operator) have been receiving considerable attention in the technical literature [1-4]. Most of this work focus on its superior performance under finite wordlength conditions when compared with those formulated with the shift-operator (or, q-operator). In particular, investigations of coefficient sensitivity and quantization noise properties have revealed that δ-operator formulations usually perform significantly better than their q-operator counterparts [1-4]. This is especially true for high-speed applications where the sampling rate is much larger than the underlying system bandwidth. Under these conditions, q-operator formulated discrete-time systems tend to become ill-conditioned [1-2].

Although a large amount of work is available on the effects of coefficientsensitivity and quantization noise, a deterministic study of the nonlinear behavior of discrete-time systems formulated with the δ-operator has not been undertaken. In the case of floating-point (FLP) arithmetic, some results for feedback system are available in [2].

In this work, we focus on the convergence behavior of the unforced system response and global asymptotic stability of δ-operator formulated discrete-time systems implemented in fixed-point (FXP) arithmetic. In particular, via necessary conditions for stability, it will be shown that such systems tend to produce DC limit cycles.

The structure of this article is as follows: In Section II, we introduce notation and nomenclature. The model for δ-operator formulated discrete-time systems, with and without quantization nonlinearities, is briefly discussed. Section III addresses the problem of asymptotic stability when FXP arithmetic is used for the implementation. In terms of ensuing DC limit cycles, necessary conditions for global asymptotic stability are formulated. It is shown that, when FXP arithmetic is used, stability of the linear system is often lost. Section IV provides concluding remarks.

II. NOTATION AND NOMENCLATURE

Since our focus is on investigation of stability properties of δ-operator formulated discrete-time systems under unforced conditions, the state equations of the system under zero-input will be considered.

In the linear case, the general m-th order state-space representation is given by

$$\delta[\mathbf{x}](n) = A^{\delta}\mathbf{x}(n); \tag{1}$$

$$\mathbf{x}(n+1) = \mathbf{x}(n) + \Delta \cdot \delta[\mathbf{x}](n), \tag{2}$$

where $\mathbf{x}(n) = [x_1(n), \ldots, x_m(n)]^T$ is the state vector at instant n, $A^{\delta} = \{a_{ij}^{\delta}\} \in \Re^{m \times m}$ is the system matrix,

and $\Delta > 0$ is the sampling time. Moreover, $\delta[\cdot]$ represents the δ-operator, that is,

$$\delta[x_\nu](n) = \frac{x_\nu(n+1) - x_\nu(n)}{\Delta}, \ \forall \nu = 1, \ldots, m, \quad (3)$$

and $\delta[\mathrm{x}](n) = [\delta[x_1](n), \ldots, \delta[x_m](n)]^T$. The actual implementation of (1) and (2) in FXP format gives rise to nonlinear quantization operations that occur at various locations depending on the hardware realization.

Eqn. (1) can be implemented either by using single wordlength accumulators (creating a quantization error after each multiplication) or by using double wordlength accumulators (creating a quantization error only after summation). We will only consider the latter option since practically all modern DSP machines implement this. Eqn. (1) can then be written as

$$\delta[\mathrm{x}](n) = Q\{A^\delta \mathrm{x}(n)\}, \quad (4)$$

where Q is a vector-valued quantization nonlinearity of the form

$$Q\{\mathrm{x}\} = \begin{pmatrix} Q\{x_1\} \\ \vdots \\ Q\{x_m\} \end{pmatrix}. \quad (5)$$

Here, $Q\{x_\nu\}$ denotes magnitude truncation, two's complement truncation, or rounding.

Eqn. (2) can be implemented in two different ways:

$$\mathrm{x}(n+1) = \mathrm{x}(n) + Q\{\Delta \cdot \delta[\mathrm{x}](n)\}, \quad (6)$$

or

$$\mathrm{x} = Q\{\mathrm{x}(n) + \Delta \cdot \delta[\mathrm{x}](n)\}. \quad (7)$$

Eqn. (6) corresponds to quantization after multiplication while (7) corresponds to quantization after summation. In contrast to (1), for (2), it is not clear which of the two quantization schemes in (6) and (7) is preferable. We will therefore consider both possibilities.

Throughout this paper, we will use the following definition of stability:

Definition. The discrete-time system in $\{(4),(6)\}$ or $\{(4),(7)\}$ is globally asymptotically stable if and only if, for any initial condition $\mathrm{x}(0)$, the state vector $\mathbf{x}$ asymptotically reaches zero, that is, $\mathrm{x}(n) \to 0$ for $n \to \infty$.

Comment. Since the FXP systems considered are in fact finite state machines, the condition $\mathrm{x}(n) \to 0$ for $n \to \infty$ may be restated as $\mathrm{x}(N) = 0$ for some finite N [5].

Finally, the symbol ℓ is used to denote the quantization step.

III. NECESSARY CONDITIONS FOR STABILITY

First, we will consider the system described by $\{(4),(6)\}$. From the definition for global asymptotic stability as stated in the previous section, it is necessary that

$$Q\{\Delta \cdot \delta[\mathrm{x}](n)\} \neq 0, \quad \text{for any} \quad \mathrm{x}(n) \neq 0. \quad (8)$$

This is just one of a finite set of conditions that is required to ensure global asymptotic stability of a FXP implementation of a linearly stable system [5].

In the case of rounding, condition (8) is violated if

$$|\Delta \cdot \delta[x_\nu](n)| \leq \frac{\ell}{2}, \quad \text{for any} \quad \nu = 1, \ldots, m. \quad (9)$$

The sampling time Δ in a δ-operator formulated implementation is typically very small. With $\Delta = I \cdot \ell$ and (9), we have

$$|\delta[x_\nu](n)| \leq \frac{1}{2I}, \quad \text{for any} \quad \nu = 1, \ldots, m, \quad (10)$$

where I is a positive integer.

In the case of magnitude truncation, (10) takes the form

$$|\delta[x_\nu](n)| \leq \frac{1}{I}, \quad \text{for any} \quad \nu = 1, \ldots m. \quad (11)$$

Accordingly, for two's complement truncation, we have

$$0 \leq \delta[x_\nu](n) < \frac{1}{I}, \quad \text{for any} \quad \nu = 1, \ldots, m. \quad (12)$$

Conditions (10-12) describe the deadband, in terms of $\delta[\mathrm{x}]$, for which a DC limit cycle occurs. Such a limit cycle can be avoided if (10-12) are satisfied by the zero vector only. In the case of rounding, we therefore require

$$\ell > \frac{1}{2I},$$

or, equivalently,

$$\Delta > \frac{1}{2}, \quad (13)$$

which is impractical. Similarly, for magnitude and two's complement truncation, we obtain

$$\ell > \frac{1}{I} \Longleftrightarrow \Delta > 1, \quad (14)$$

which again is equally impractical.

This result is summarized in the following theorem.

Theorem 1. A necessary condition for stability of the δ-operator formulated discrete-time system in $\{(4),(6)\}$ is $\Delta > 0.5$ for rounding and $\Delta > 1$ for truncation.

The above theorem shows that high-speed δ-operator formulated implementations that possess a small sampling time cannot be realized limit cycle free in FXP format!

A second necessary condition for the system in $\{(4),(6)\}$ can be obtained by noting that

$$\delta[\mathbf{x}](n) = \mathbf{0} \tag{15}$$

can occur in (4) even though the state vector $\mathbf{x}(n) \neq \mathbf{0}$.

Therefore, for rounding, no nonzero state vector $\mathbf{x}(n)$ that satisfies

$$-\begin{pmatrix} \frac{\ell}{2} \\ \vdots \\ \frac{\ell}{2} \end{pmatrix} \leq A^{\delta} \cdot \mathbf{x}(n) \leq +\begin{pmatrix} \frac{\ell}{2} \\ \vdots \\ \frac{\ell}{2} \end{pmatrix} \tag{16}$$

may be allowed to exist. Here, the inequality has to hold elementwise. Taking norms on both sides of (16) one gets an algebraic condition on the system matrix A^{δ} that always support DC limit cycles. Eqn. (16) has the following interesting interpretations:

1. Each of the resulting m inequalities can be geometrically interpreted as the intersection of two half spaces in $\Re^m$. These intersections are symmetric about the origin and have parallel boundaries. The normal vector to the boundaries is given by the particular row vector of A^{δ}. Only if the intersection of *all* such m half spaces contains a nonzero point in $\Re^m$, and if it belongs to the quantization lattice, will there exist a nonzero state vector that is an equilibrium point of the system.

2. Eqn. (16) can also be interpreted from an eigenvalue/eigenvector viewpoint. In high-speed digital filters where the sampling frequency is typically much higher than the bandwidth of the processed signal, a q-operator implementation's eigenvalues cluster around the point $z = 1$ [1]. The corresponding δ-operator implementation for large sampling times has eigenvalues clustered around zero. However, as the sampling time becomes small, these eigenvalues move towards the eigenvalues of the underlying continuous-time system [1]. In other words, for large sampling times, the system matrix will be ill-conditioned, that is, vectors $\mathbf{x}(n) \neq 0$ exist such that $A^{\delta} \cdot \mathbf{x}(n)$ is close to the zero vector. According to (16), this is likely to cause a DC limit cycle. For small sampling times, this problem may not occur; however, in this case, the conditions in Theorem 1 are not satisfied!

In the case of the remaining two quantization schemes, the inequalities corresponding to (16) are given as follows: For two's complement truncation,

$$\mathbf{0} \leq A^{\delta} \cdot \mathbf{x}(n) < \begin{pmatrix} \ell \\ \vdots \\ \ell \end{pmatrix}, \ \mathbf{x}(n) \neq \mathbf{0}, \tag{17}$$

and, for magnitude truncation,

$$-\begin{pmatrix} \ell \\ \vdots \\ \ell \end{pmatrix} < A^{\delta} \cdot \mathbf{x}(n) < +\begin{pmatrix} \ell \\ \vdots \\ \ell \end{pmatrix}, \ \mathbf{x}(n) \neq \mathbf{0}. \tag{18}$$

A similar analysis can be conducted for the system in $\{(4),(7)\}$. Since (4) is common to both realizations, (16-18) are still valid and provide conditions under which the finite difference is quantized to zero and a DC limit cycle is produced. We will now briefly discuss necessary conditions for global asymptotic stability obtained from (7).

For rounding, proceeding as in (9), we have

$$|\Delta \cdot \delta[x_\nu](n)| \leq \frac{\ell}{2}, \quad \text{for any} \quad \nu = 1, \ldots, m,$$

and therefore

$$|\delta[x_\nu](n)| \leq \frac{1}{2I}, \quad \text{for any} \quad \nu = 1, \ldots, m. \tag{19}$$

For magnitude truncation, we obtain

$$0 \leq \delta[x_\nu](n) < \frac{1}{I}, \ \forall \delta[x_\nu] \geq 0, \tag{20}$$

and

$$-\frac{1}{I} < \delta[x_\nu](n) \leq 0, \ \forall \delta[x_\nu] < 0. \tag{21}$$

In the case of two's complement truncation, the condition for a DC limit cycle is given by

$$0 \leq \delta[x_\nu](n) < \frac{1}{I}, \ \forall \nu = 1, \ldots, m. \tag{22}$$

With $\Delta = I \cdot \ell$, I being a 'small' integer, we come to the same conclusion as for the previously considered system:

$$\Delta > \frac{1}{2} \quad \text{for rounding;}$$

$$\Delta > 1 \quad \text{for truncation.}$$

Therefore, Theorem 1 also holds for the system representation in $\{(4),(7)\}$.

IV. CONCLUSION

Via a set of necessary conditions for global asymptotic stability, it has been shown that high-speed, limit cycle free δ-operator implementations of linear discrete-time systems cannot be realized. This is due to the tendency of such a realization to produce period one limit cycles. This situation arises from small values in the finite difference being quantized to zero. Hence, convergence to the 'wrong' equilibrium point is very likely. Conditions on the system matrix and the sampling time if such limit cycle behavior is to be avoided have been provided. The results indicate that, in high-speed applications, these conditions cannot be satisfied with conventional quantization schemes.

ACKNOWLEDGEMENT

This work was partially supported by a research grant from the Office of Naval Research (ONR).

REFERENCES

[1] G.C. Goodwin, R.H. Middleton, and H.V. Poor, "High speed digital signal processing and control," *Proceedings of the IEEE*, 80, 2, pp. 240-259, Feb. 1992.

[2] R.H. Middleton and G.C. Goodwin, "Improved finite wordlength characteristics in digital control using delta-operators," *IEEE Transactions Automatic Control*, 31, 11, pp. 1015-1021, Nov. 1986.

[3] G. Li and M. Gevers, "Comparative study of finite wordlength effects in shift and delta operator parameterization," *Proceedings of the 1990 IEEE Conference on Decision and Control (CDC'90)*, 2, pp. 954-959, Honolulu, HI, Dec. 1990.

[4] G. Li and M. Gevers, "Roundoff noise minimization using delta-operator realizations," *IEEE Transactions on Signal Processing*, 41, 2, pp. 629-637, Feb. 1993.

[5] P.H. Bauer and L.J. Leclerc, "A computer-aided test for the absence of limit cycles in fixed point digital filters," *IEEE Transactions on Signal Processing*, 39, 11, pp. 2400-2410, Nov. 1991.

[6] K. Premaratne, R. Salvi, N.R. Habib, and J.P. LeGall, "Delta-operator formulated discrete-time approximations of continuous-time systems," to appear in *IEEE Transactions on Automatic Control*, 1994.

Proceedings of the 37th Midwest Symposium on Circuits and Systems (MWSCAS'94), Lafayette, LA, Aug. 03-05, 1994, pp. 805-808.

Two-Dimensional Delta-Operator Formulated Discrete-Time Systems: State-Space Realization and Its Coefficient Sensitivity Properties

K. Premaratne, J. Suarez, and M.M. Ekanayake
Department of E&CE
University of Miami
Coral Gables, FL 33124 USA

P.H. Bauer
Department of EE
Laboratory of Image and Signal Analysis
University of Notre Dame
Notre Dame, IN 46556 USA

Abstract—By developing the δ-operator analog of the Roesser model, state-space realization of two- and multi-dimensional δ-systems is investigated. The corresponding notions of gramians and balanced realization are also defined. It is shown that, discrete-time system implementation using this model can yield superior coefficient sensitivity properties.

I. Introduction

Judging by its performance in the one-dimensional (1-D) case [2], [5-6], one is led to expect superior coefficient sensitivity and roundoff noise performance with δ-operator implementation of two-dimensional (2-D) and multi-dimensional (m-D) discrete-time (DT) systems. With this in mind, δ-operator analog of the q-operator Roesser local state-space (s.s.) model [12] is developed. We also propose the notions of gramians and balanced (BL) realization. As expected, realization using this model can provide superior coefficient sensitivity properties.

II. Nomenclature and Preliminaries

A. Nomenclature

$\Re$: Reals; $\Im$: Complex numbers; $\Re^{q\times p}$, $\Im^{q\times p}$: Matrices of size $q \times p$ over $\Re$ and $\Im$; I_n: $n \times n$ unit matrix; A^*, trace$[A]$, $\|A\|_F$: Conjugate transpose, trace, and Fröbenius norm of matrix A; $\mathrm{e}_i^{(n)}$: Unit vector in $\Re^n$ with 1 on the i-th row; $E_{i,j}^{q\times p} = \mathrm{e}_i^{(q)}\mathrm{e}_j^{(p)^*} \in \Re^{q\times p}$; $\overline{U}_{q\times p} = \sum_{i=1}^{q}\sum_{j=1}^{p} E_{i,j}^{(q\times p)} \otimes E_{i,j}^{(q\times p)} \in \Re^{q^2\times p^2}$.

For q- and δ-systems, we use the indeterminates z and c, respectively. For 1-D systems, $\delta = (q-1)/\tau \Longleftrightarrow c = (z-1)/\tau$, where τ is a positive real constant, usually the sampling time. Let $\overline{\mathcal{U}}_\delta^2 = \{(c_h, c_v) \in \Im^2 : |c_h + 1/\tau_h| \le 1/\tau_h, |c_v + 1/\tau_v| \le 1\}$. $\mathcal{T}_\delta^2$ is its boundary. The corresponding q-system regions are denoted with the subscript q.

K.P. and P.H.B. gratefully acknowledge the support received from the Office of Naval Research (ONR) through the grants N00014-94-1-0454 and N00014-94-1-0387, respectively.

B. Preliminaries

Consider a linear, shift-invariant, strictly causal, p-input q-output 2-D DT system. Its n_hh-n_vv Roesser local s.s. model $\{\hat{A}, \hat{B}, \hat{C}, \hat{D}\}$ takes the form [12]:

$$\begin{aligned}\begin{bmatrix} q_h[\mathbf{x}^h](i,j) \\ q_v[\mathbf{x}^v](i,j)\end{bmatrix} &= [\hat{A}]\begin{bmatrix}\mathbf{x}^h(i,j)\\ \mathbf{x}^v(i,j)\end{bmatrix} + [\hat{B}]\mathbf{u}(i,j);\\ \mathbf{y}(i,j) &\doteq [\hat{C}]\begin{bmatrix}\mathbf{x}^h(i,j)\\ \mathbf{x}^v(i,j)\end{bmatrix} + [\hat{D}]\mathbf{u}(i,j),\end{aligned} \tag{2.1}$$

where $\mathbf{u} \in \Re^p$, $\mathbf{x}^h \in \Re^{n_h}$, $\mathbf{x}^v \in \Re^{n_v}$, and $\mathbf{y} \in \Re^q$. $\mathbf{x}^h$ and $\mathbf{x}^v$ are the h.p. and v.p. local state vectors. Take $n = n_h + n_v$. Also,

$$q_h[\mathbf{x}](i,j) = \mathbf{x}(i+1,j); \quad q_v[\mathbf{x}](i,j) = \mathbf{x}(i,j+1). \tag{2.2}$$

In what follows, we use matrix partitioning that conform to $A \doteq \begin{bmatrix}\hat{A}^{(1)} & \hat{A}^{(2)}\\ \hat{A}^{(3)} & \hat{A}^{(4)}\end{bmatrix}$, $B \doteq \begin{bmatrix}\hat{B}^{(1)}\\ \hat{B}^{(2)}\end{bmatrix}$, and $C \doteq \begin{bmatrix}\hat{C}^{(1)}, \hat{C}^{(2)}\end{bmatrix}$. The corresponding 2-D characteristic equation and transfer function are

$$\begin{aligned}\det[I_z - \hat{A}] &= \det[z_h I_{n_h} \oplus z_v I_{n_v} - \hat{A}];\\ \hat{H}(z_h, z_v) &= \hat{C}(I_z - \hat{A})^{-1}\hat{B} + \hat{D},\end{aligned} \tag{2.3}$$

where $z_h, z_v \in \Im$ $I_z \doteq z_h I_{n_h} \oplus z_v I_{n_v} \in \Im^{n\times n}$. With no nonessential singularities of the second kind (NSSK) on $\mathcal{T}_q^2$, $\{\hat{A}, \hat{B}, \hat{C}, \hat{D}\}$ is BIBO stable iff [3]

$$\det[I_z - \hat{A}] \neq 0, \ \forall (z_h, z_v) \in \overline{\mathcal{U}}_q^2. \tag{2.4}$$

III. 2-D δ-Model

A. Local s.s. model

Analogous to the 1-D case, define $\delta_h[\cdot]$ and $\delta_v[\cdot]$ as

$$\begin{aligned}\delta_h[\mathbf{x}](i,j) &= \frac{\mathbf{x}(i+1,j) - \mathbf{x}(i,j)}{\tau_h} = \frac{q_h[\mathbf{x}](i,j) - \mathbf{x}(i,j)}{\tau_h};\\ \delta_v[\mathbf{x}](i,j) &= \frac{\mathbf{x}(i,j+1) - \mathbf{x}(i,j)}{\tau_v} = \frac{q_v[\mathbf{x}](i,j) - \mathbf{x}(i,j)}{\tau_v}.\end{aligned} \tag{3.1}$$

Here τ_h and τ_v are positive real constants denoting the 'sampling times' along h.p. and v.p. directions, respectively. Note that

$$q_h = 1 + \tau_h \delta_h; \quad q_v = 1 + \tau_v \delta_v, \tag{3.2}$$

and letting $\tau = [\tau_h I_{n_h} \oplus \tau_v I_{n_v}] \in \Re^{n \times n}$,

$$\begin{bmatrix} q_h[\mathbf{x}^h](i,j) \\ q_v[\mathbf{x}^v](i,j) \end{bmatrix} = I_n + \tau \begin{bmatrix} \delta_h I_{n_h} & 0 \\ 0 & \delta_v I_{n_v} \end{bmatrix} \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix}. \tag{3.3}$$

Using (3.3) in (2.1), we get

$$\begin{aligned} \begin{bmatrix} \delta_h[\mathbf{x}^h](i,j) \\ \delta_v[\mathbf{x}^v](i,j) \end{bmatrix} &\doteq [A] \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} + [B]\mathbf{u}(i,j); \\ \mathbf{y}(i,j) &\doteq [C] \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} + [D]\mathbf{u}(i,j), \end{aligned} \tag{3.4}$$

where $A \doteq \begin{bmatrix} A^{(1)} & A^{(2)} \\ A^{(3)} & A^{(4)} \end{bmatrix}$, $B \doteq \begin{bmatrix} B^{(1)} \\ B^{(2)} \end{bmatrix}$, and $C \doteq \left[C^{(1)}, C^{(2)}\right]$. In addition, we need to perform

$$q_h[\mathbf{x}^h] = \mathbf{x}^h + \tau_h \cdot \delta_h[\mathbf{x}^h]; \; q_v[\mathbf{x}^v] = \mathbf{x}^v + \tau_v \cdot \delta_v[\mathbf{x}^v]. \tag{3.5}$$

Here,

$$\hat{A} = I_n + \tau A; \; \hat{B} = \tau B; \; \hat{C} = C; \; \hat{D} = D. \tag{3.6}$$

B. Properties of the 2-D δ-model

Most of the following properties may be derived in a manner that is exactly analogous to that in [12].

The *transition matrix* $A^{i,j}$ of the δ-model, may be recursively computed from

$$A^{i,j} = \begin{cases} 0, (i,j) = (0,0); \\ [I_{n_h} \oplus I_{n_v}], (i,j) = (0,0); \\ \begin{bmatrix} I_{n_h} & 0 \\ 0 & 0 \end{bmatrix} + \tau \begin{bmatrix} A^{(1)} & A^{(2)} \\ 0 & 0 \end{bmatrix}, (i,j) = (1,0); \\ \begin{bmatrix} 0 & 0 \\ 0 & I_{n_v} \end{bmatrix} + \tau \begin{bmatrix} 0 & 0 \\ A^{(3)} & A^{(4)} \end{bmatrix}, (i,j) = (0,1); \\ A^{1,0} A^{i-1,j} + A^{0,1} A^{i,j-1}, \text{ elsewhere.} \end{cases} \tag{3.7}$$

The general response of the δ-model is

$$\begin{aligned} \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} &= \sum_{k=0}^{j} A^{i,j-k} \begin{bmatrix} \mathbf{x}^h(0,k) \\ 0 \end{bmatrix} \\ &\quad + \sum_{h=0}^{i} A^{i-j,j} \begin{bmatrix} 0 \\ \mathbf{x}^v(h,0) \end{bmatrix} + \mathbf{f}(\mathbf{u}), \end{aligned} \tag{3.8}$$

where $\mathbf{f}(\mathbf{u}) = \sum_{(0,0) \le (h,k) < (i,j)} (A^{i-h-1,j-k} \tau \begin{bmatrix} B^{(1)} \\ 0 \end{bmatrix} + A^{i-h,j-k-1} \tau \begin{bmatrix} 0 \\ B^{(2)} \end{bmatrix}) \mathbf{u}(h,k)$.

Let $I_c \doteq c_h I_{n_h} \oplus c_v I_{n_v} \in \Im^{n \times n}$. Then, the 2-D δ-model's characteristic equation and transfer function are

$$\begin{aligned} \det[I_c - A] &= \frac{1}{\det[\tau]} \det[I_z - \hat{A}]|_{\mathbf{z} \to \mathbf{c}}; \\ H(c_h, c_v) &= \hat{H}(z_h, z_v)|_{\mathbf{z} \to \mathbf{c}}, \end{aligned} \tag{3.9}$$

where

$$z_h = 1 + \tau_h c_h; \quad z_v = 1 + \tau_v c_v. \tag{3.10}$$

From now on, the variable transformation in (3.10) is denoted by $\mathbf{c} \to \mathbf{z}$ or $\mathbf{z} \to \mathbf{c}$ whatever is appropriate.

Nonsingular transformations of the type

$$\begin{bmatrix} \tilde{\mathbf{x}}^h(i,j) \\ \tilde{\mathbf{x}}^v(i,j) \end{bmatrix} = [T] \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix}, \tag{3.11}$$

where $T \doteq [T^{(1)} \oplus T^{(4)}]$, yield the equivalent 2-D s.s. realization $\{\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}\}$. Here,

$$\tilde{A} = TAT^{-1}; \; \tilde{B} = TB; \; \tilde{C} = CT^{-1}; \; \tilde{D} = D. \tag{3.12}$$

With no NSSK on $\mathcal{T}_\delta^2$, $\{A, B, C, D\}$ is BIBO stable iff

$$\det[I_c - A] \neq 0, \; \forall (c_h, c_v) \in \overline{U}_\delta^2. \tag{3.13}$$

C. Gramians

The gramians of 2-D q-systems are taken to be natural extensions of the integral expressions of their 1-D counterparts [11]. We will adopt a similar approach. In what follows, we consider the 1-D (or 2-D) stable δ-system $\{A, B, C, D\}$ with gramians P and Q. The corresponding q-system is $\{\hat{A}, \hat{B}, \hat{C}, \hat{D}\}$ with gramians $\hat{P}$ and $\hat{Q}$.

1-D case. The gramians are defined in [10].

Definition 3.1. [10]. The gramians are the solutions to the Lyapunov equations

$$\begin{aligned} AP + PA^* + \tau \cdot APA^* &= -BB^*; \\ A^*Q + QA + \tau \cdot A^*QA &= -C^*C. \end{aligned}$$

Lemma 3.1. The gramians satisfy the integral expressions

$$P = \frac{1}{2\pi j} \oint_{\mathcal{T}_q} FF^* \frac{dc}{1 + \tau c}; \; Q = \frac{1}{2\pi j} \oint_{\mathcal{T}_q} G^*G \frac{dc}{1 + \tau c},$$

where $F(c) \doteq (cI_n - A)^{-1} B$ and $G(c) \doteq C(cI_n - A)^{-1}$. Moreover, $\hat{P} = \tau P$ and $\hat{Q} = Q/\tau$.

Proof. Substitute $\hat{A} = I_n + \tau A$, $\hat{B} = \tau B$, $\hat{C} = C$, and $\hat{D} = D$ [10] in the equations in Definition 3.1, and note the integral expressions for P and Q in [8]. ■

2-D case. With Lemma 3.1 in mind, we have

Definition 3.2. The gramians are

$$P = \frac{1}{(2\pi j)^2} \oint_{\mathcal{T}_\delta^2} FF^* \frac{dc_h}{1+\tau_h c_h} \frac{dc_v}{1+\tau_v c_v};$$

$$Q = \frac{1}{(2\pi j)^2} \oint_{\mathcal{T}_\delta^2} G^*G \frac{dc_h}{1+\tau_h c_h} \frac{dc_v}{1+\tau_v c_v},$$

where $P \doteq \begin{bmatrix} P^{(1)} & P^{(2)} \\ P^{(3)} & P^{(4)} \end{bmatrix}$ and $Q \doteq \begin{bmatrix} Q^{(1)} & Q^{(2)} \\ Q^{(3)} & Q^{(4)} \end{bmatrix}$. Also, $F(c_h, c_v) \doteq (I_c - A)^{-1}B = [\mathrm{f}_1, \ldots, \mathrm{f}_n]^*$ and $G(c_h, c_v) \doteq C(I_c - A)^{-1} = [\mathrm{g}_1, \ldots, \mathrm{g}_n]$.

Remarks.

1. Note that, $(I_c - A)^{-1}|_{\mathbf{c}\to\mathbf{z}} = (I_z - \hat{A})^{-1}\tau$, and

$$F|_{\mathbf{c}\to\mathbf{z}} = \hat{F};\ G|_{\mathbf{c}\to\mathbf{z}} = \hat{G}\cdot\tau. \tag{3.14}$$

2. Definition 3.2 is completely analogous to the 1-D and 2-D q-systems [7], [11].

Lemma 3.2. $\hat{P} = \tau_h\tau_v P$ and $\hat{Q} = \tau_h\tau_v \tau^{-1} Q \tau^{-1}$.

Proof. Consider P in Definition 3.2. Use $\mathbf{c} \to \mathbf{z}$, (3.14), and definition of gramians for 2-D q-systems [11]. ■

The following are in complete analogy with 2-D q-systems.

Lemma 3.3. The gramians may be represented as

$$P = \frac{1}{\tau_h\tau_v} \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} M_{i,j} M_{ij}^*;$$

$$Q = \frac{1}{\tau_h\tau_v} \tau \cdot \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} A^{i,j^*} C^* C A^{i,j} \cdot \tau,$$

where, for $(i,j) = (0,0)$, $M_{i,j} = 0$, and, for $(i,j) > (0,0)$, $M_{ij} \doteq A^{i-1,j}\tau \begin{bmatrix} B^{(1)} \\ 0 \end{bmatrix} + A^{i,j-1}\tau \begin{bmatrix} 0 \\ B^{(2)} \end{bmatrix}$.

Lemma 3.4. Let $\{\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}\}$ with gramians $\tilde{P}$ and $\tilde{Q}$ be an equivalent system as in (3.10-11). Then, $\tilde{P} = TPT^*$ and $\tilde{Q} = T^{-1^*} Q T^{-1}$. Moreover, the eigenvalues of PQ and $\tilde{P}\tilde{Q}$ are invariant.

Definition 3.3. $\{A, B, C, D\}$ is said to be *balanced* if $P^{(1)} = Q^{(1)} \doteq \Sigma^{(1)} = \mathrm{diag}\{\sigma_1^{(1)}, \sigma_2^{(1)}, \ldots, \sigma_{n_h}^{(1)}\}$ and $P^{(4)} = Q^{(4)} \doteq \Sigma^{(4)} = \mathrm{diag}\{\sigma_1^{(4)}, \sigma_2^{(4)}, \ldots, \sigma_{n_v}^{(4)}\}$.

If the diagonal submatrices of P and Q are each positive definite (p.d.), a BL realization may be obtained [4]. Regarding this, we have

Lemma 3.5. Local reachability and observability of $\{A, B, C, D\}$ and $\{\hat{A}, \hat{B}, \hat{C}, \hat{D}\}$ are equivalent. Moreover, when $\{A, B, C, D\}$ is locally reachable and observable, $P^{(1)}$, $P^{(4)}$, $Q^{(1)}$, and $Q^{(4)}$ are each p.d.

Separable systems. A separable (in denominator) 2-D q-system will have $\hat{A}^{(2)} = 0$ (and/or $\hat{A}^{(3)} = 0$) and all off-diagonal submatrices of $\hat{P}$ and $\hat{Q}$ are zero. The diagonal submatrices may be computed through two pairs of Lyapunov equations [11]. Clearly, a separable 2-D q-system yields a separable 2-D δ-system.

Theorem 3.6. Let $\{A, B, C, D\}$ be separable with $A^{(2)} = 0$. Then, $P^{(2)} = Q^{(2)} = 0$ and $P^{(3)} = Q^{(3)} = 0$, and

$$\begin{aligned}
&A^{(1)}P^{(1)} + P^{(1)}A^{(1)^*} + \tau_h A^{(1)} P^{(1)} A^{(1)^*} \\
&\quad = -B^{(1)}B^{(1)^*}/\tau_v; \\
&A^{(1)^*}Q^{(1)} + Q^{(1)}A^{(1)} + \tau_h A^{(1)^*} Q^{(1)} A^{(1)} \\
&\quad = -\begin{bmatrix} C^{(1)} & R^{(4)}A^{(3)} \end{bmatrix}^* \begin{bmatrix} C^{(1)} & R^{(4)}A^{(3)} \end{bmatrix} / \tau_v; \\
&A^{(4)}P^{(4)} + P^{(4)}A^{(4)^*} + \tau_v A^{(4)} P^{(4)} A^{(4)^*} \\
&\quad = -\begin{bmatrix} B^{(2)} & A^{(3)}S^{(1)} \end{bmatrix} \begin{bmatrix} B^{(2)} & A^{(3)}S^{(1)} \end{bmatrix}^* / \tau_h; \\
&A^{(4)^*}Q^{(4)} + Q^{(4)}A^{(4)} + \tau_v A^{(4)^*} Q^{(4)} A^{(4)} \\
&\quad = -C^{(2)^*}C^{(2)}/\tau_h.
\end{aligned}$$

Here, $R^{(4)^*}R^{(4)} \doteq \tau_h\tau_v Q^{(4)}$ and $S^{(1)}S^{(1)^*} \doteq \tau_h\tau_v P^{(1)}$.

IV. Coefficient Sensitivity

By generalizing a certain sensitivity measure, Lutz and Hakimi [9] have addressed sensitivity minimization of MIMO 1-D CT systems. The SISO 2-D q-operator case is in [7]. In what follows, we study the coefficient sensitivity of the 2-D δ-model in section III. We follow a more direct approach using Kronecker product formulation and, hence, the results are applicable to the more general MIMO case. Using [1], we may show

$$S_A(c_h, c_v) = [I_n \otimes G] \cdot \overline{U}_{n\times n} \cdot [I_n \otimes F] \tag{4.1}$$

$$S_B(c_h, c_v) = [I_n \otimes G] \cdot \overline{U}_{n\times p} \tag{4.2}$$

$$S_C(c_h, c_v) = \overline{U}_{q\times n} \cdot [I_n \otimes F] \tag{4.3}$$

$$S_D(c_h, c_v) = \overline{U}_{q\times p} \tag{4.4}$$

Lemma 4.1. The quantities in (4.1-4.4) are given as

$$S_A = \begin{bmatrix} \mathrm{g}_1 \\ \vdots \\ \mathrm{g}_n \end{bmatrix} [\mathrm{f}_1^* \quad \cdots \quad \mathrm{f}_n^*];$$

$$S_B = \begin{bmatrix} \mathrm{g}_1^{(1)} & \cdots & \mathrm{g}_1^{(p)} \\ \vdots & \ddots & \vdots \\ \mathrm{g}_n^{(1)} & \cdots & \mathrm{g}_n^{(p)} \end{bmatrix};$$

$$S_C = \begin{bmatrix} \mathrm{f}_1^{(1)^*} & \cdots & \mathrm{f}_n^{(1)^*} \\ \vdots & \ddots & \vdots \\ \mathrm{f}_1^{(q)^*} & \cdots & \mathrm{f}_n^{(q)^*} \end{bmatrix};$$

$$S_D = \begin{bmatrix} E_{1,1} & \cdots & E_{1,p} \\ \vdots & \ddots & \vdots \\ E_{q,1} & \cdots & E_{q,p} \end{bmatrix}.$$

Here, $\mathbf{f}_i^{(j)^*}$ denotes a $(q \times p)$ null matrix except its j-th row which is $\mathbf{f}_i^*$ and $\mathbf{g}_i^{(j)}$ denotes a $(q \times p)$ null matrix except its j-th column which is $\mathbf{g}_i$.

Proof. This may be shown through the results in [1] and simple yet tedious algebraic manipulations. ■

Corollary 4.2. The quantities S_A, S_B, S_C, and S_D of the δ-model and the quantities $\hat{S}_{\hat{A}}$, $\hat{S}_{\hat{B}}$, $\hat{S}_{\hat{C}}$, and $\hat{S}_{\hat{D}}$ of the corresponding q-model are related by $S_A|_{\mathbf{c}\to\mathbf{z}} = \tau\hat{S}_{\hat{A}}$, $S_B|_{\mathbf{c}\to\mathbf{z}} = \tau\hat{S}_{\hat{B}}$, $S_C|_{\mathbf{c}\to\mathbf{z}} = \hat{S}_{\hat{C}}$, and $S_D|_{\mathbf{c}\to\mathbf{z}} = \hat{S}_{\hat{D}}$, where $\tau = \tau_h I_{n_h q} \oplus \tau_v I_{n_v q} \in \Re^{nq \times nq}$.

Proof. Apply (3.14) to Lemma 4.1. ■

To proceed further, we utilize the following

Definition 4.1. Let $H(c_h, c_v)$ be a bivariate matrix-valued function that is analytic on $\mathcal{T}_\delta^2$. Then,

$$\|H(c_h, c_v)\|_p^p \doteq \frac{1}{(2\pi\jmath)^2} \oint_{\mathcal{T}_q^2} \|H(c_h, c_v)|_{\mathbf{c}\to\mathbf{z}}\|_F^p \frac{dz_h}{z_h} \frac{dz_v}{z_v}.$$

Remark. This norm is extensively utilized in related work [7] due mainly to the fact that it leads to tractable results. This, and our desire to make a comparison with the corresponding q-model, are the primary reasons for its use here.

We now define the absolute sensitivity measure

$$M \doteq \|S_A\|_1^2 + \frac{1}{p}\|S_B\|_2^2 + \frac{1}{q}\|S_C\|_2^2 + \frac{1}{pq}\|S_D\|_2^2. \quad (4.5)$$

Remarks.

1. The use of different norms is for mathematical feasibility and tractability [7], [5].
2. The weights associated with each term in (4.5) may be thought of as *averaging factors per input/output.*
3. Due to (3.5), M should contain $\|S_{\tau_h}\|$ and $\|S_{\tau_v}\|$. However, we assume that τ_h and τ_v are selected such that each possess exact binary representations. Hence, these additional terms are neglected.

Using an argument similar to that in [7], one may show the following:

$$\|S_A\|_1^2 \le \text{trace}[\hat{P}] \cdot \text{trace}[\tau\hat{Q}\tau] \quad (4.6)$$

$$\|S_B\|_2^2 = p \cdot \text{trace}[\tau\hat{Q}\tau] \quad (4.7)$$

$$\|S_c\|_2^2 = q \cdot \text{trace}[\hat{P}] \quad (4.8)$$

$$\|S_D\|_2^2 = pq \quad (4.9)$$

Combining (4.5) with (4.6-9), we get

$$M \le \overline{M} \doteq (\text{trace}[\hat{P}] + 1)(\text{trace}[\tau\hat{Q}\tau] + 1). \quad (4.10)$$

It is customary to perform a minimization of $\overline{M}$. Hence, one attempts to characterize those $\{\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}\}$ that are 'bound optimal' with respect to M. Analogous to 2-D q-systems case [7], one may for instance show that a BL realization (modulo an orthogonal nonsingular transformation) is 'bound optimal' with respect to M.

Compared to a q-system, its δ-system counterpart yields a smaller $\overline{M}$ whenever $\text{trace}[\hat{Q}] > \text{trace}[\tau\hat{Q}\tau]$, that is,

$$(1 - \tau_h^2) \cdot \text{trace}[\hat{Q}^{(1)}] + (1 - \tau_v^2) \cdot \text{trace}[\hat{Q}^{(4)}] > 0. \quad (4.11)$$

Note that, with the local reachability and observability assumption of $\{A, B, C, D\}$, p.d. of $Q^{(1)}$ and $Q^{(4)}$ (and hence of $\hat{Q}^{(1)}$ and $\hat{Q}^{(4)}$) are guaranteed. Thus, (4.11) is satisfied if $\tau_h < 1$ and $\tau_v < 1$.

VII. Conclusion

We have developed the δ-operator analog of the Roesser local s.s. model. Notions of gramians and BL realization are also proposed. As is expected, under mild conditions, this model offers superior coefficient sensitivity properties.

References

[1] J.W. Brewer, "Kronecker products and matrix calculus in system theory," *IEEE Trans. Circ. Syst.*, vol. CAS-25, pp. 772-781, Sept. 1978.

[2] G.C. Goodwin, R.H. Middleton, and H.V. Poor, "High-speed digital signal processing and control," *Proc. IEEE*, vol. 80, pp. 240-259, 1992.

[3] E.I. Jury, "Stability of multidimensional systems and other related problems," in *Multidimensional Systems, Techniques, and Applications*, S.G. Tzafestas, Ed., New York: Marcel Dekkar, 1986.

[4] A.J. Laub, M.T. Heath, C.C. Paige, and R.C. Ward, "Computation of system balancing transformations and other applications of simultaneous diagonalization algorithms," *IEEE Trans. Auto. Cont.*, vol. AC-32, pp. 115-122, Feb. 1987.

[5] G. Li and M. Gevers, "Comparative study of finite wordlength effects in shift and delta operator parameterizations," *Proc. CDC'90*, Honolulu, Dec. 1990, pp. 954-959.

[6] G. Li and M. Gevers, "Roundoff noise minimization using delta-operator realizations," *IEEE Trans. Sig. Proc.*, vol. 41, pp. 629-637, Feb. 1993.

[7] T. Lin, M. Kawamata, and T. Higuchi, "Minimization of sensitivity of 2-D systems and its relation to 2-D balanced realizations," *Proc. ISCAS'87*, Philadelphia, May 1987, vol. 2, pp. 710-713.

[8] W.S. Lu, E.B. Lee, and Q.T. Zhang, "Model reduction for two-dimensional systems," *Proc. ISCAS'86*, 1986, vol. 1, pp. 79-82.

[9] W.J. Lutz and S.L. Hakimi, "Design of multi-input multi-output systems with minimum sensitivity," *IEEE Trans. Circ. Syst.*, vol. 35, pp. 1114-1122, Sept. 1988.

[10] R.H. Middleton and G.C. Goodwin, *Digital Control and Estimation: A Unified Approach*, Englewood Cliffs: Prentice-Hall, 1990.

[11] K. Premaratne, E.I. Jury, and M. Mansour, "An algorithm for model reduction of 2-D discrete-time systems," *IEEE Trans. Circ. Syst.*, vol. CAS-37, pp. 1116-1132, Sept. 1990.

[12] R.P. Roesser, "A discrete state model for linear image processing," *IEEE Trans. Auto. Cont.*, vol. AC-20, pp. 1-10, Feb. 1975.

Proceedings of the SOUTHCON'95, Fort Lauderdale, FL, to appear, March 1995

On Balanced Realizations of 2-D Delta-Operator Formulated Discrete-Time Systems

K. Premaratne and M.M. Ekanayake
Department of Electrical and
Computer Engineering
University of Miami
Coral Gables, FL 33124 USA
kprema@umiami.ir.miami.edu

P.H. Bauer
Department of Electrical Engineering
Laboratory of Image and Signal Analysis
University of Notre Dame
Notre Dame, IN 46556 USA
pbauer@mars.ee.nd.edu

ABSTRACT

Delta-operator based implementations can avoid the numerical ill-conditioning usually associated with high speed shift-operator based implementations of discrete-time systems. Moreover, it provides a unified methodology for tackling both continuous- and discrete-time systems. In particular, it has been shown that, delta-operator based balanced realizations can offer superior coefficient sensitivity properties under fixed-point arithmetic. In this work, we address computation of balanced realizations. For this purpose, given a discrete-time system, the relationship between its shift- and delta-operator formulated balanced realizations is presented.

I. INTRODUCTION

Current interest in delta-systems (δ-systems) is due mainly to two reasons: (a) δ-systems provide superior roundoff noise [1-2] and coefficient sensitivity [3-4] properties, and (b) δ-operator makes it possible to treat both continuous-time (CT) and discrete-time (DT) systems in a unified manner [5]. Recent work on δ-operator based implementation of two-dimensional (2-D) DT systems contain the counterpart to the shift-operator (q-operator) based Roesser local state-space (s.s.) model [6]. Balanced (BL) realization of such models and coefficient sensitivity properties were also investigated. Indeed, given a 2-D DT system, under fixed-point (FXP) arithmetic (and mild conditions), Roesser δ-model was shown to be superior to the Roesser q-model. In this paper, we reveal the relationship between BL realizations of Roesser δ- and q-models. This makes it possible to use techniques available for computation of q-BL models for computation of δ-BL models.

II. NOMENCLATURE AND PRELIMINARIES

2.1. Nomenclature

$\Re$, $\Im$, and $\aleph$ denote the reals, complex numbers, and non-negative integers, respectively. $\Re^{q\times p}$ and $\Im^{q\times p}$ are the sets of matrices of size $q \times p$ over $\Re$ and $\Im$, respectively.

I_n is the unit matrix of size $n \times n$; $\mathbf{0}$ is the null matrix of size $q \times p$. A^* and A^T denote the complex conjugate transpose and transpose of matrix $A \in \Re^{q\times p}$; trace$[A]$ and $\lambda_i[A]$ denote its trace and i-th eigenvalue. $\|A\|_F$ is its Fröbenius norm.

In the 1-D case, corresponding q- and δ-systems are related by $\delta = (q-1)/\Delta \iff c = (z-1)/\Delta$. Here, Δ is a positive real constant (usually the sampling time). For 2-D systems, subscripts h and v denote horizontally propagating (h.p.) and vertically propagating (v.p.) subsystems of the corresponding Roesser local s.s. models. n_h and n_v denote the sizes of these h.p. and v.p. subsystems. We use n to denote $n = n_h + n_v$. Δ_h and Δ_v are positive real constants denoting 'sampling times' along h.p. and v.p. directions.

We use ξ to denote $\Delta_h I_{n_h} \oplus \Delta_v I_{n_v} \in \Re^{n\times n}$. Also, I_z and I_c denote $z_h I_{n_h} \oplus z_v I_{n_v} \in \Im^{n\times n}$ and $c_h I_{n_h} \oplus c_v I_{n_v} \in \Im^{n\times n}$, respectively.

Corresponding 2-D q- and δ-systems are related by $\delta_h = (q_h - 1)/\Delta_h \iff c_h = (z_h - 1)/\Delta_h$ and $\delta_v = (q_v - 1)/\Delta_v \iff c_v = (z_v - 1)/\Delta_v$. We use subscripts δ and q to differentiate between corresponding δ- and q-systems; for example, s.s. realization of a given DT system is either $\{A_\delta, B_\delta, C_\delta, D_\delta\}$ if implemented based on δ-operator or $\{A_q, B_q, C_q, D_q\}$ if implemented based on q-operator. The following notation is also used: $H(c_h, c_v)|_{\mathbf{c}\to\mathbf{z}} = H(c_h, c_v)|_{\substack{c_h=(z_h-1)/\Delta_h \\ c_v=(z_v-1)/\Delta_v}}$ and $G(z_h, z_v)|_{\mathbf{z}\to\mathbf{c}} = G(z_h, z_v)|_{\substack{z_h=1+\Delta_h c_h \\ z_v=1+\Delta_v c_v}}$.

Stability studies of q- and δ-systems involve the follow-

ing regions: $\mathcal{U}_q = \{z \in \Im : |z| < 1\}$; $\mathcal{U}_q^2 = \{(z_h, z_v) \in \Im^2 : |z_h| < 1, |z_v| < 1\}$; $\mathcal{U}_\delta = \{c \in \Im : |c + 1/\Delta| < 1/\Delta\}$; $\mathcal{U}_\delta^2 = \{(c_h, c_v) \in \Im^2 : |c_h + 1/\Delta_h| < 1/\Delta_h, |c_v + 1/\Delta_v| < 1\}$. The corresponding distinguished boundaries are denoted with letter $\mathcal{T}$.; $\mathcal{U}. \cup \mathcal{T}.$ is denoted by $\overline{\mathcal{U}}.$. A q-system polynomial with all its roots in $\mathcal{U}_q$ (for the 1-D case) or $\mathcal{U}_q^2$ (for the 2-D case) is said to be *stable*. The corresponding regions for a δ-system polynomial are $\mathcal{U}_\delta$ (for the 1-D case) and $\mathcal{U}_\delta^2$ (for the 2-D case), respectively.

2.2. Preliminaries

First, we provide a brief overview of relevant material.

Roesser q-model. The 2-D dynamical system under consideration is assumed to be linear, shift-invariant, strictly causal, and modeled by a set of first-order vector difference equations over $\Re$. Given such a p-input and q-output system, its n_hh-n_vv Roesser local s.s. model $\{A_q, B_q, C_q, D_q\}$ is of the form [7]

$$\begin{aligned} \begin{matrix} q_h[\mathbf{x}^h](i,j) \\ q_v[\mathbf{x}^v](i,j) \end{matrix} &= \begin{bmatrix} A_q^{(1)} & A_q^{(2)} \\ A_q^{(3)} & A_q^{(4)} \end{bmatrix} \begin{matrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{matrix} + \begin{bmatrix} B_q^{(1)} \\ B_q^{(2)} \end{bmatrix} \mathbf{u}(i,j) \\ &\doteq [A_q] \begin{matrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{matrix} + [B_q]\mathbf{u}(i,j); \\ \mathbf{y}(i,j) &= \begin{bmatrix} C_q^{(1)} & C_q^{(2)} \end{bmatrix} \begin{matrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{matrix} + [D_q]\mathbf{u}(i,j) \\ &\doteq [C_q] \begin{matrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{matrix} + [D_q]\mathbf{u}(i,j), \end{aligned} \tag{2.1}$$

where $\mathbf{u} \in \Re^p$, $\mathbf{x}^h \in \Re^{n_h}$, $\mathbf{x}^v \in \Re^{n_v}$, and $\mathbf{y} \in \Re^q$. Also, $A_q^{(1)} \in \Re^{n_h \times n_h}$, $A_q^{(4)} \in \Re^{n_v \times n_v}$, etc. Here, $(i,j) \in \aleph^2$ and

$$q_h[\mathbf{x}](i,j) = \mathbf{x}(i+1,j) \quad \text{and} \quad q_v[\mathbf{x}](i,j) = \mathbf{x}(i,j+1). \tag{2.2}$$

Usually, $\mathbf{x}^h$ and $\mathbf{x}^v$ are called the *h.p.* and *v.p.* local state vectors of $\{A_q, B_q, C_q, D_q\}$. With no nonessential singularities of the second kind on $\mathcal{T}_q^2$, for BIBO stability, one requires [8]

$$\det[I_z - A_q] \neq 0, \ \forall (z_h, z_v) \in \overline{\mathcal{U}}_q^2. \tag{2.3}$$

Roesser δ-model. To exploit the superior finite wordlength properties of δ-operator implementations, analogous to the 1-D case, in [6], the following operators are defined:

$$\begin{aligned} \delta_h[\mathbf{x}](i,j) &= \frac{\mathbf{x}(i+1,j) - \mathbf{x}(i,j)}{\Delta_h} = \frac{q_h[\mathbf{x}](i,j) - \mathbf{x}(i,j)}{\Delta_h}; \\ \delta_v[\mathbf{x}](i,j) &= \frac{\mathbf{x}(i,j+1) - \mathbf{x}(i,j)}{\Delta_v} = \frac{q_v[\mathbf{x}](i,j) - \mathbf{x}(i,j)}{\Delta_v}, \end{aligned} \tag{2.4}$$

where Δ_h and Δ_v are two positive real numbers. Hence, the following relationships are applicable:

$$\delta_h = \frac{q_h - 1}{\Delta_h}; \qquad \delta_v = \frac{q_v - 1}{\Delta_v}. \tag{2.5}$$

Using (2.4-5) in (2.1), the following Roesser δ-model $\{A_\delta, B_\delta, C_\delta, D_\delta\}$ has been proposed [6]:

$$\begin{aligned} \begin{matrix} \delta_h[\mathbf{x}^h](i,j) \\ \delta_v[\mathbf{x}^v](i,j) \end{matrix} &= \begin{bmatrix} A_\delta^{(1)} & A_\delta^{(2)} \\ A_\delta^{(3)} & A_\delta^{(4)} \end{bmatrix} \begin{matrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{matrix} + \begin{bmatrix} B_\delta^{(1)} \\ B_\delta^{(2)} \end{bmatrix} \mathbf{u}(i,j) \\ &\doteq [A_\delta] \begin{matrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{matrix} + [B_\delta]\mathbf{u}(i,j); \\ \mathbf{y}(i,j) &= \begin{bmatrix} C_\delta^{(1)} & C_\delta^{(2)} \end{bmatrix} \begin{matrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{matrix} + [D_\delta]\mathbf{u}(i,j) \\ &\doteq [C_\delta] \begin{matrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{matrix} + [D_\delta]\mathbf{u}(i,j), \end{aligned} \tag{2.6}$$

where

$$A_\delta = \xi^{-1}(A_q - I_n); \ B_\delta = \xi^{-1}B_q; \ C_\delta = C_q; \ D_\delta = D_q. \tag{2.7}$$

Here, $\xi = [\Delta_h I_{n_h} \oplus \Delta_v I_{n_v}] \in \Re^{n \times n}$. Note that, as opposed to its corresponding Roesser q-model, here, one must also perform the following computations:

$$q_h[\mathbf{x}^h] = \mathbf{x}^h + \Delta_h \cdot \delta_h[\mathbf{x}^h]; \quad q_v[\mathbf{x}^v](i,j) = \mathbf{x}^v + \Delta_v \cdot \delta_v[\mathbf{x}^v]. \tag{2.8}$$

In [6], several properties of this Roesser δ-model (such as, general response equation, transition matrix, characteristic equation, transfer function) are elaborated. Also, it is easy to see that, as for the q-model, 2-D equivalent transformations of the type

$$\begin{bmatrix} \tilde{\mathbf{x}}^h(i,j) \\ \tilde{\mathbf{x}}^v(i,j) \end{bmatrix} = \begin{bmatrix} T^{(1)} & 0 \\ 0 & T^{(4)} \end{bmatrix} \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} \doteq [T] \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix}, \tag{2.9}$$

where $T^{(1)} \in \Re^{n_h \times n_h}$ and $T^{(4)} \in \Re^{n_v \times n_v}$ are nonsingular, yield an equivalent 2-D s.s. realization $\{\tilde{A}_\delta, \tilde{B}_\delta, \tilde{C}_\delta, \tilde{D}_\delta\}$, where

$$\tilde{A}_\delta = TA_\delta T^{-1}; \ \tilde{B}_\delta = TB_\delta; \ \tilde{C}_\delta = C_\delta T^{-1}; \ \tilde{D}_\delta = D_\delta. \tag{2.10}$$

Also, $\{\tilde{A}_\delta, \tilde{B}_\delta, \tilde{C}_\delta, \tilde{D}_\delta\}$ and $\{A_\delta, B_\delta, C_\delta, D_\delta\}$ have the same transfer function. With no nonessential singularities of the second kind on $\mathcal{T}_\delta^2$, for BIBO stability, one requires

$$\det[I_c - A_\delta] \neq 0, \ \forall (c_h, c_v) \in \overline{\mathcal{U}}_\delta^2. \tag{2.11}$$

III. GRAMIANS AND BL REALIZATIONS

3.1. Gramians

For the Roesser q-model, gramians are taken to be natural extensions of the integral expressions of their 1-D counterparts [9-10]. The work in [6] adopts a similar approach in proposing gramians for the δ-operator case as defined in [5]. In what follows, $\{A_\delta, B_\delta, C_\delta, D_\delta\}$ (with gramians P_δ and Q_δ) and $\{A_q, B_q, C_q, D_q\}$ (with gramians P_q and Q_q) denote a given stable 2-D DT system's δ- and q-operator based Roesser models, respectively.

DEFINITION 3.1. [9-10].

1. Gramians of $\{A_q, B_q, C_q, D_q\}$ are

$$\begin{aligned} P_q &= \frac{1}{(2\pi j)^2} \oint \oint_{\mathcal{T}_q^2} F_q F_q^* \frac{dz_h}{z_h} \frac{dz_v}{z_v}; \\ Q_q &= \frac{1}{(2\pi j)^2} \oint \oint_{\mathcal{T}_q^2} G_q^* G_q \frac{dz_h}{z_h} \frac{dz_v}{z_v}, \end{aligned}$$

where $F_q(z_h, z_v) = (I_z - A_q)^{-1}B_q$ and $G_q(z_h, z_v) = C_q(I_z - A_q)^{-1}$.

2. Gramians of $\{A_\delta, B_\delta, C_\delta, D_\delta\}$ are

$$P_\delta = \frac{1}{(2\pi j)^2}\oint_{T_\delta^2} F_\delta F_\delta^* \frac{dc_h}{1+\Delta_h c_h}\frac{dc_v}{1+\Delta_v c_v};$$

$$Q_d d = \frac{1}{(2\pi j)^2}\oint_{T_\delta^2} G_\delta^* G_\delta \frac{dc_h}{1+\Delta_h c_h}\frac{dc_v}{1+\Delta_v c_v},$$

where $F_\delta(c_h, c_v) = (I_c - A_\delta)^{-1}B_\delta$ and $G_\delta(c_h, c_v) = C_\delta(I_c - A_\delta)^{-1}$.

LEMMA 3.1. [6]. The relationship between the above gramians are

$$P_\delta = \frac{1}{\Delta_h\Delta_v}P_q; \qquad Q_\delta = \frac{1}{\Delta_h\Delta_v}\xi Q_q \xi.$$

With appropriate partitions incorporated, this is equivalent to

$$\begin{bmatrix} P_\delta^{(1)} & P_\delta^{(2)} \\ P_\delta^{(3)} & P_\delta^{(4)} \end{bmatrix} = \frac{1}{\Delta_h\Delta_v}\begin{bmatrix} P_q^{(1)} & P_q^{(2)} \\ P_q^{(3)} & P_q^{(4)} \end{bmatrix};$$

$$\begin{bmatrix} Q_\delta^{(1)} & Q_\delta^{(2)} \\ Q_\delta^{(3)} & Q_\delta^{(4)} \end{bmatrix} = \begin{bmatrix} \frac{\Delta_h}{\Delta_v}Q_q^{(1)} & Q_q^{(2)} \\ Q_q^{(3)} & \frac{\Delta_v}{\Delta_h}Q_q^{(4)} \end{bmatrix}.$$

LEMMA 3.2. [6]. The realization $\{\tilde{A}_\delta, \tilde{B}_\delta, \tilde{C}_\delta, \tilde{D}_\delta\}$ obtained with a nonsingular transformation of the type in (2.9-10) yields the gramians $\tilde{P}_\delta = TP_\delta T^*$ and $\tilde{Q}_\delta = T^{-1^*}Q_\delta T^{-1}$. Eigenvalues of $P_\delta Q_\delta$ are invariant under such a transformation. The situation regarding Roesser q-model is completely equivalent.

DEFINITION 3.2. [10]. Roesser δ-model $\{A_\delta, B_\delta, C_\delta, D_\delta\}$ is said to be *balanced (BL)* if

$$P_\delta^{(1)} = Q_\delta^{(1)} \doteq \Sigma_\delta^{(1)} = \mathrm{diag}\{\sigma_{\delta_1}^{(1)}, \ldots, \sigma_{\delta_{n_h}}^{(1)}\};$$

$$P_\delta^{(4)} = Q_\delta^{(4)} \doteq \Sigma_\delta^{(4)} = \mathrm{diag}\{\sigma_{\delta_1}^{(4)}, \ldots, \sigma_{\delta_{n_v}}^{(4)}\}.$$

We refer to $\sigma_{\delta_i}^{(1)}$, $i = 1, \ldots, n_h$, and $\sigma_{\delta_j}^{(4)}$, $j = 1, \ldots, n_v$, as the *Hankel singular values* of h.p. and v.p. subsystems, repectively. The situation regarding Roesser q-model is completely equivalent.

3.2. Computation of BL Realizations

Computation of gramians and obtaining BL realizations for q-systems have been investigated quite thoroughly. In the 1-D and 2-D separable cases, one may solve Lyapunov equations and use Laub's algorithm [10-11]. In the 2-D nonseparable case, this computation is not that easy; however, several techniques have been developed [10], [12].

In this section, we provide the relationship between BL realizations of corresponding δ- and q-models. This allows all available techniques for gramian computation of q-systems to be utilized for δ-systems as well. To the authors' knowledge, such a relationship is not available even for the 1-D case. Although we concentrate on the 2-D case, a similar argument may be developed for the 1-D case.

For convenience, we use the following notation:
$\{A, B, C, D\} \xrightarrow{T} \{\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}\}$: Here, $\tilde{A} = TAT^{-1}$, $\tilde{B} = TB$, $\tilde{C} = CT^{-1}$, and $\tilde{D} = D$, where T is of type (2.9-10).
$\{A_q, B_q, C_q, D_q\} \xrightarrow{q\to\delta} \{A_\delta, B_\delta, C_\delta, D_\delta\}$: This is the corresponding δ-system obtained by applying (2.7).
$\{A_\delta, B_\delta, C_\delta, D_\delta\} \xrightarrow{\delta\to q} \{A_q, B_q, C_q, D_q\}$: This is the corresponding q-system obtained by applying (2.7).

Moreover, we use the following:
$\{A_{qB}, B_{qB}, C_{qB}, D_{qB}\}$: BL realization of $\{A_q, B_q, C_q, D_q\}$ obtained via $\{A_q, B_q, C_q, D_q\} \xrightarrow{T_q} \{A_{qB}, B_{qB}, C_{qB}, D_{qB}\}$.
$\{A_{\delta B}, B_{\delta B}, C_{\delta B}, D_{\delta B}\}$: BL realization of $\{A_\delta, B_\delta, C_\delta, D_\delta\}$ obtained via $\{A_\delta, B_\delta, C_\delta, D_\delta\} \xrightarrow{T_\delta} \{A_{\delta B}, B_{\delta B}, C_{\delta B}, D_{\delta B}\}$.
$\{A_{\delta B2q}, B_{\delta B2q}, C_{\delta B2q}, D_{\delta B2q}\}$: q-system obtained via $\{A_{\delta B}, B_{\delta B}, C_{\delta B}, D_{\delta B}\} \xrightarrow{\delta\to q} \{A_{\delta B2q}, B_{\delta B2q}, C_{\delta B2q}, D_{\delta B2q}\}$.
$\{A_{qB2\delta}, B_{qB2\delta}, C_{qB2\delta}, D_{qB2\delta}\}$: δ-system obtained via $\{A_{qB}, B_{qB}, C_{qB}, D_{qB}\} \xrightarrow{q\to\delta} \{A_{qB2\delta}, B_{qB2\delta}, C_{qB2\delta}, D_{qB2\delta}\}$.

LEMMA 3.3. The following relationships are true:

$$\{A_q, B_q, C_q, D_q\} \xrightarrow{T_\delta} \{A_{\delta B2q}, B_{\delta B2q}, C_{\delta B2q}, D_{\delta B2q}\};$$

$$\{A_\delta, B_\delta, C_\delta, D_\delta\} \xrightarrow{T_q} \{A_{qB2\delta}, B_{qB2\delta}, C_{qB2\delta}, D_{qB2\delta}\}.$$

Proof. Note that, $A_{\delta B2q} = I_n + \xi A_{\delta B} = I_n + \xi T_\delta A_\delta T_\delta^{-1} = I_n + \xi T_\delta \xi^{-1}(A_q - I_n)T_\delta^{-1} = T_\delta A_q T_\delta^{-1}$, since $\xi T_\delta \xi^{-1} = T_\delta$. The remainder may proven in a similar manner. ∎

LEMMA 3.4. The following relationships are true:

$$\{A_{\delta B2q}, B_{\delta B2q}, C_{\delta B2q}, D_{\delta B2q}\} \xrightarrow{\xi^{-1/2}} \{A_{qB}, B_{qB}, C_{qB}, D_{qB}\};$$

$$\{A_{qB2\delta}, B_{qB2\delta}, C_{qB2\delta}, D_{qB2\delta}\} \xrightarrow{\xi^{1/2}} \{A_{\delta B}, B_{\delta B}, C_{\delta B}, D_{\delta B}\}.$$

Proof. Note that, $\{A_{\delta B}, B_{\delta B}, C_{\delta B}, D_{\delta B}\}$ has following gramians:

$$P_{\delta B} = \begin{bmatrix} \Sigma_\delta^{(1)} & P_{\delta B}^{(2)} \\ P_{\delta B}^{(3)} & \Sigma_\delta^{(4)} \end{bmatrix}; \quad Q_{\delta B} = \begin{bmatrix} \Sigma_\delta^{(1)} & Q_{\delta B}^{(2)} \\ Q_{\delta B}^{(3)} & \Sigma_\delta^{(4)} \end{bmatrix}.$$

Hence, from Lemma 3.1, $\{A_{\delta B2q}, B_{\delta B2q}, C_{\delta B2q}, D_{\delta B2q}\}$ has the following gramians:

$$P_{\delta B2q} = \Delta_h\Delta_v \begin{bmatrix} \Sigma_\delta^{(1)} & P_{\delta B}^{(2)} \\ P_{\delta B}^{(3)} & \Sigma_\delta^{(4)} \end{bmatrix};$$

$$Q_{\delta B2q} = \begin{bmatrix} \frac{\Delta_v}{\Delta_h}\Sigma_\delta^{(1)} & Q_{\delta B}^{(2)} \\ Q_{\delta B}^{(3)} & \frac{\Delta_h}{\Delta_v}\Sigma_\delta^{(4)} \end{bmatrix}.$$

To get $\{A_{qB}, B_{qB}, C_{qB}, D_{qB}\}$, we need to simultaneously diagonalize the two pairs $\{\Delta_h\Delta_v\Sigma_\delta^{(1)}, (\Delta_v/\Delta_h)\Sigma_\delta^{(1)}\}$ and $\{\Delta_h\Delta_v\Sigma_\delta^{(4)}, (\Delta_h/\Delta_v)\Sigma_\delta^{(4)}\}$. By applying Laub's algorithm, we get these two transformations to be $\Delta_h^{-1/2}I_{n_h}$ and

$\Delta_v^{-1/2} I_{n_v}$. This proves the first part. The remainder follows in a similar manner. ∎

COROLLARY 3.5. The systems $\{A_{qB}, B_{qB}, C_{qB}, D_{qB}\}$ and $\{A_{\delta B}, B_{\delta B}, C_{\delta B}, D_{\delta B}\}$ are related as follows:

$$A_{\delta B} = \xi^{-1/2}(A_{qB} - I_n)\xi^{-1/2};\ B_{\delta B} = \xi^{-1/2} B_{qB};$$
$$C_{\delta B} = C_{qB}\xi^{-1/2};\ D_{\delta B} = D_{qB}.$$

Proof. Note that, from Lemma 3.4, $A_{\delta B} = \xi^{-1}(A_{\delta B 2q} - I_n) = \xi^{-1}(\xi^{1/2} A_{qB} \xi^{-1/2} - I_n) = \xi^{-1/2}(A_{qB} - I_n)\xi^{-1/2}$. The rest follows in a similar manner. ∎

IV. EXAMPLE

We now consider a stable 3h-3v 2-D separable digital filter.

4.1. Computations

Numerical values are *displayed* via FORMAT SHORT E of MATLAB [13] which was used for all computations. Note that, since system being considered has $A_q^{(3)} = \mathbf{0}$ (instead of $A_q^{(2)} = \mathbf{0}$), relevant equations must be appropriately modified.

Given q-model $\{A_q, B_q, C_q, D_q\}$.

$$A_q^{(1)} = \begin{bmatrix} 0 & 1.0000e+00 & 0 \\ 0 & 0 & 1.0000e+00 \\ 3.8315e-01 & -1.3861e+00 & 1.9067e+00 \end{bmatrix};$$

$$A_q^{(2)} = \begin{bmatrix} -6.8280e-02 & 6.1900e-02 & 6.5400e-03 \\ -2.8100e-02 & 3.9560e-02 & -2.2480e-02 \\ 1.2445e+00 & -5.7092e-01 & 2.0587e+00 \end{bmatrix};$$

$$A_q^{(3)} = \mathbf{0}; \text{ or } \begin{bmatrix} 0 & 1.0000e+00 & 0 \\ A_q^{(4)} \ 0 & 0 & 1.0000e+00 \\ 3.8238e-01 & -1.3818e+00 & 1.9025e+00 \end{bmatrix};$$

$$B_q^{(1)} = [0 \quad 0 \quad 1]^T;$$
$$B_q^{(2)} = [0 \quad 0 \quad 1]^T;$$
$$C_q^{(1)} = [1.1410e-02 \quad -5.4000e-03 \quad 1.9560e-02];$$
$$C_q^{(2)} = [1.1640e-02 \quad -5.4500e-03 \quad 1.9600e-02];$$
$$D_q = [9.4300e-03].$$

BL q-model $\{A_{qB}, B_{qB}, C_{qB}, D_{qB}\}$.

$$A_{qB}^{(1)} = \begin{bmatrix} 8.6478e-01 & 2.6806e-01 & -3.4799e-02 \\ -2.6806e-01 & 5.8766e-01 & 3.8402e-01 \\ -3.4797e-02 & -3.8401e-01 & 4.5427e-01 \end{bmatrix};$$

$$A_{qB}^{(2)} = \begin{bmatrix} 4.2940e-01 & -3.3765e-01 & 1.2689e-01 \\ 3.3771e-01 & -2.6511e-01 & 1.0134e-01 \\ 1.2732e-01 & -9.7518e-02 & 3.2423e-02 \end{bmatrix};$$

$$A_{qB}^{(3)} = \mathbf{0};$$

$$A_{qB}^{(4)} = \begin{bmatrix} 8.6486e-01 & 2.6760e-01 & -3.4949e-02 \\ -2.6760e-01 & 5.8692e-01 & 3.8661e-01 \\ -3.4952e-02 & -3.8661e-01 & 4.5071e-01 \end{bmatrix};$$

$$B_{qB}^{(1)} = [6.3568e-02 \quad 4.9879e-02 \quad 1.8565e-02]^T;$$
$$B_{qB}^{(2)} = [6.5595e-01 \quad 5.1555e-01 \quad 1.9416e-01];$$
$$C_{qB}^{(1)} = [6.5590e-01 \quad -5.1574e-01 \quad 1.9341e-01];$$
$$C_{qB}^{(2)} = [6.3592e-02 \quad -4.9875e-02 \quad 1.8540e-02];$$
$$D_{qB} = [9.4300e-03].$$

Corresponding δ-model $\{A_\delta, B_\delta, C_\delta, D_\delta\}$. Let us select $\Delta_h = Delta_v = 2.5000e-01$. Acoordingly, we get

$$A_\delta^{(1)} = \begin{bmatrix} -4.0000e+00 & 4.0000e+00 & 0 \\ 0 & -4.0000e+00 & 4.0000e+00 \\ 1.5326e+00 & -5.5444e+00 & 3.6268e+00 \end{bmatrix};$$

$$A_\delta^{(2)} = \begin{bmatrix} -2.7312e-01 & 2.4760e-01 & 2.6160e-02 \\ -1.1240e-01 & 1.5824e-01 & -8.9920e-02 \\ 4.9780e+00 & -2.2837e+00 & 8.2348e+00 \end{bmatrix};$$

$$A_\delta^{(3)} = \mathbf{0};$$

$$A_\delta^{(4)} = \begin{bmatrix} -4.0000e+00 & 4.0000e+00 & 0 \\ 0 & -4.0000e+00 & 4.0000e+00 \\ 1.5295e+00 & -5.5272e+00 & 3.6100e+00 \end{bmatrix};$$

$$B_\delta^{(1)} = [0 \quad 0 \quad 4]^T;$$
$$B_\delta^{(2)} = [0 \quad 0 \quad 4]^T;$$
$$C_\delta^{(1)} = [1.1410e-02 \quad -5.4000e-03 \quad 1.9560e-02];$$
$$C_\delta^{(2)} = [1.1640e-02 \quad -5.4500e-03 \quad 1.9600e-02];$$
$$D_\delta = [9.4300e-03].$$

BL δ-model $\{A_{\delta B}, B_{\delta B}, C_{\delta B}, D_{\delta B}\}$.

$$A_{\delta B}^{(1)} = \begin{bmatrix} -5.4089e-01 & 1.0722e+00 & -1.3919e-01 \\ -1.0722e+00 & -1.6494e+00 & 1.5361e+00 \\ -1.3919e-01 & -1.5361e+00 & -2.1829e+00 \end{bmatrix};$$

$$A_{\delta B}^{(2)} = \begin{bmatrix} 1.7176e+00 & -1.3506e+00 & 5.0755e-01 \\ 1.3508e+00 & -1.0604e+00 & 4.0537e-01 \\ 5.0926e-01 & -3.9007e-01 & 1.2969e-01 \end{bmatrix};$$

$$A_{\delta B}^{(3)} = \mathbf{0};$$

$$A_{\delta B}^{(4)} = \begin{bmatrix} -5.4054e-01 & 1.0704e+00 & -1.3980e-01 \\ -1.0704e+00 & -1.6523e+00 & 1.5464e+00 \\ -1.3981e-01 & -1.5464e+00 & -2.1971e+00 \end{bmatrix};$$

$$B_{\delta B}^{(1)} = [1.2714e-01 \quad 9.9759e-02 \quad 3.7129e-02]^T;$$
$$B_{\delta B}^{(2)} = [1.3119e+00 \quad 1.0311e+00 \quad 3.8833e-01]^T;$$
$$C_{\delta B}^{(1)} = [1.3118e+00 \quad -1.0315e+00 \quad 3.8682e-01];$$
$$C_{\delta B}^{(2)} = [1.2718e-01 \quad -9.9750e-02 \quad 3.7080e-02];$$
$$D_{\delta B} = [9.4300e-03].$$

4.2. Simulations

Normalized frequency response of $\{A_q, B_q, C_q, D_q\}$ is $H_q(e^{j\omega_1}, e^{j\omega_2})$ and that of $\{A_\delta, B_\delta, C_\delta, D_\delta\}$ is $H_\delta((e^{j\omega_1}-1)/\Delta_h, (e^{j\omega_2}-1)/\Delta_v)$. Frequency responses are evaluated on $\mathcal{G}^2 \doteq \{(\omega_1, \omega_2) \in \Re^2 : \omega_i = n_i \times \pi/N,\ n_i = [-N : 1 : N],\ i = 1, 2\}$ with $N = 32$. For comparison purposes, the following measure was also evaluated: For $(z_1, z_2) = (e^{j\omega_1}, e^{j\omega_2})$ and $(c_1, c_2) = ((e^{j\omega_1}-1)/\Delta_h, (e^{j\omega_2}-1)/\Delta_v)$,

$$E_{\max} \doteq \begin{cases} \max_{\mathcal{G}^2} \left| H(z_1, z_2) - \hat{H}(z_1, z_2) \right|, & \text{for } q\text{-models;} \\ \max_{\mathcal{G}^2} \left| H(c_1, c_2) - \hat{H}(c_1, c_2) \right|, & \text{for } \delta\text{-models.} \end{cases}$$

Here, H denotes the 'ideal' frequency response where each coefficient is represented in 'infinite' precision; $\hat{H}$ denotes the 'actual' frequency response where each coefficient is represented in finite precision.

Fig. (1) shows $E_{\max}$ versus number of fractional bits where each coefficient is represented in FXP and its fractional part is truncated at different lengths; integral part is represented exactly. Advantage gained by BL δ-model over BL q-model is about 3 bits.

Fig. (2) shows $E_{\max}$ versus total number of bits where each coefficient is represented in FXP and its total (integral+fractional) number of bits is truncated at different lengths. Advantage gained by BL δ-model over BL q-model is only about 1 bit. This modest improvement is due to the large Δ_h and Δ_v being used. More dramatic improvements require smaller Δ_h and Δ_v [6]. But, this makes δ-model's coefficients to occupy a larger dynamic range. To circumvent this, we believe, careful scaling of filter coefficients is necessary. We are currently investigating this possibility.

Fig. (3) shows $E_{\max}$ versus number of mantissa bits where each coefficient is represented in FLP and its number of mantissa bits is truncated at different lengths. Of course, in FLP, dynamic range is usually of no threat.

V. CONCLUSION

In this work, we have presented the relationship between BL realizations of corresponding δ- and q-models. This, in turn, addresses computation of gramians and BL realizations of δ-models.

In the FXP case, δ-model is better whenever $\Delta_h < 1$ and $\Delta_v < 1$ [6]. However, this choice must be carefully done since, in FXP, δ-models tend to occupy a larger dynamic range. The authors are currently investigating the possibility of incorporating scaling of coefficients so that low values of Δ_h and Δ_v may be used to expose and exploit the advantages of δ-systems. In the FLP case, such a limitation does not usually arise, and δ-models are better whenever the system matrix eigenvalues lie within a certain region called the *MG-region* [14]. This condition is typically true for high Q, narrowband digital filters operating under high sampling rates. These observations indicate that, in FLP, for comparative performance (with respect to coefficient sensitivity), δ-models require a shorter mantissa length. The ensuing implications regarding low power consumption, low cost and weight, and high speed cannot be overemphasized. The authors are currently completing work regarding quantization noise properties of the δ-model developed, where, as in 1-D case, improvements over the corresponding q-model are expected.

We must mention that certain difficulties regarding limit cycles are inherent in δ-systems when FXP arithmetic is used [15]. However, this problem is, for all practical purposes, nonexistent in FLP arithmetic. Hence, in our opinion, for FLP high performance applications, the δ-model developed provides an extremely attractive solution that avoids numerical ill-conditioning typically associated with high speed q-systems.

REFERENCES

[1] G.C. Goodwin, R.H. Middleton, and H.V. Poor, "High-speed digital signal processing and control," *Proc. IEEE,* vol. 80, pp. 240-259, 1992.

[2] G. Li and M. Gevers, "Roundoff noise minimization using delta-operator realizations," *IEEE Trans. Sig. Proc.,* vol. 41, pp. 629-637, Feb. 1993.

[3] G. Li and M. Gevers, "Comparative study of finite wordlength effects in shift and delta operator parameterizations," *Proc. 1990 IEEE Conf. Decision and Cont. (CDC'90),* pp. 954-959, Honolulu, HI, Dec. 1990.

[4] K. Premaratne, R. Salvi, N.R. Habib, and J.P. Le Gall, "Delta-operator formulated discrete-time equivalents of continuous-time systems," *IEEE Trans. Auto. Cont.,* vol. 39, pp. 581-585, Mar. 1994.

[5] R.H. Middleton and G.C. Goodwin, *Digital Control and Estimation: A Unified Approach,* Englewood Cliffs, NJ: Prentice-Hall, 1990.

[6] K. Premaratne, J. Suarez, M.M. Ekanayake, and P.H. Bauer, "Two-dimensional delta-operator formulated discrete-time systems: State-space realization and its coefficient sensitivity properties," *Proc. 37th Midwest Symp. Circ. Syst. (MWSCS'94),* Lafayette, LA, Aug. 1994.

[7] R.P. Roesser, "A discrete state model for linear image processing," *IEEE Trans. Auto. Cont.,* vol. AC-20, pp. 1-10, Feb. 1975.

[8] E.I. Jury, "Stability of multidimensional systems and other related problems," Chapter 3 in *Multidimensional Systems, Techniques, and Applications,* New York, NY: Marcel Dekkar, 1986.

[9] W.-S. Lu, E.B. Lee, and Q.T. Zhang, "Model reduction for two-dimensional systems," *Proc. 1986 IEEE Int. Symp. Circ. Syst. (ISCAS'86),* vol. 1, pp. 79-82, 1986.

[10] K. Premaratne, E.I. Jury, and M. Mansour, "An algorithm for model reduction of 2-D discrete-time systems," *IEEE Trans. Circ. Syst.,* vol. CAS-37, pp. 1116-1132, Sept. 1990.

[11] A.J. Laub, M.T. Heath, C.C. Paige, and R.C. Ward, "Computation of system balancing transformations and other applications of simultaneous diagonalization algorithms," *IEEE Trans. Auto. Cont.,* vol. AC-32, pp. 115-122, Feb. 1987.

[12] W.-S. Lu, H.-P. Wang, and A. Antoniou, "An efficient method for the evaluation of the controllability and observability gramians of 2-D digital filters and systems," *IEEE Trans. Circ. Syst.—II. Anal. Dig. Sig. Proc.,* vol. 39, pp. 695-704, Oct. 1992.

[13] *MATLAB,* ver. 4.2a, Natick, MA: The MathWorks Inc.

[14] K. Premaratne, M.M. Ekanayake, J. Suarez, and P.H. Bauer, "Two-dimensional delta-operator formulated discrete-time systems: State-space realization and its coefficient sensitivity properties" (detailed version of [6]), *IEEE Trans. Sig. Proc.,* in review, 1995.

[15] K. Premaratne and P.H. Bauer, "Limit cycles and asymptotic stability of delta-operator systems in fixed-point arithmetic," *Proc. 1994 IEEE Int. Symp. Circ. Syst. (ISCAS'94),* London, UK, vol. 2, pp. 461-464, May 1994.

ACKNOWLEDGEMENT

The work of K.P. and P.H.B. were partially supported by the US Office of Naval Research (ONR) through grants N00014-94-1-0454 and N00014-94-1-0387, respectively.

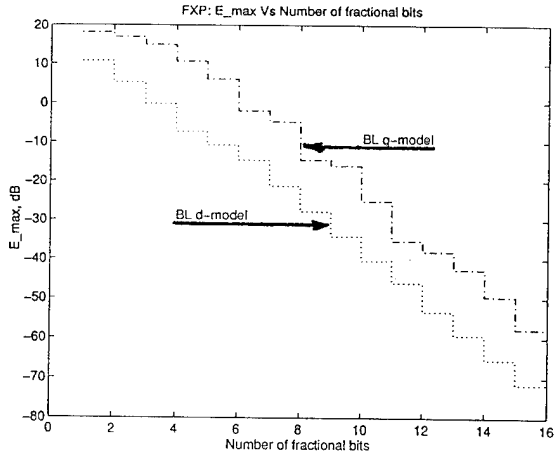


Figure (1)

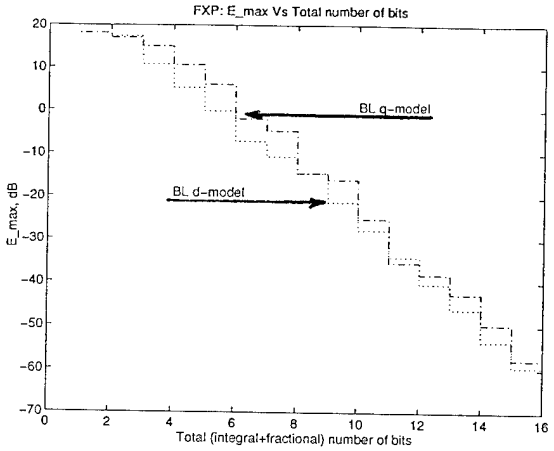


Figure (2)

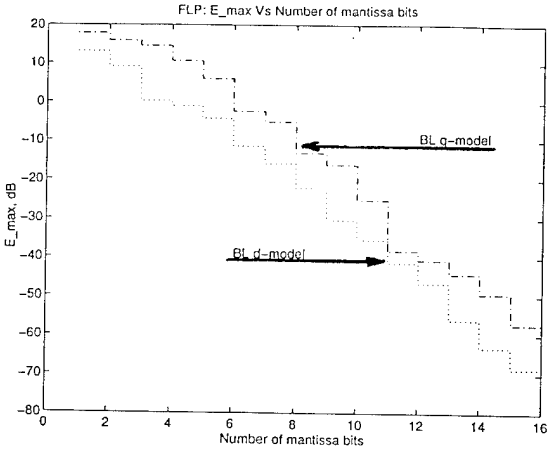


Figure (3)

1995 International Symposium on Circuits and Systems (ISCAS'95), Seattle, WA, to appear, April-May, 1995.

An Exhaustive Search Algorithm For Checking Limit Cycle Behavior Of Digital Filters

K. Premaratne, *Senior Member, IEEE*, E.C. Kulasekere
P.H. Bauer, *Member, IEEE*, L.J. Leclerc

Abstract—**In this paper, an algorithm that can be utilized to determine the presence or absence of limit cycles in fixed-point implementation of digital filters is given. It is applicable for filters in state-space formulation (and hence, application to the corresponding direct form follows as a special case), and is independent of the order, type of quantization, and whether the accumulator is single- or double-length. Bounds on the amplitude and period of possible limit cycles are presented. The robustness of the algorithm in terms of limit cycle performance with respect to filter coefficient perturbations is verified. The algorithm is then used to obtain regions in the coefficient space where a filter of given order is limit cycle free. In this process, we have obtained limit cycle free regions that were previously unknown for the Two's complement case.**

I. Introduction

In realizing a digital filter, its coefficients and intermediate results of computations must be stored in registers of finite wordlength. Care must be taken to suppress resulting limit cycles as otherwise performance degradation may render the design unacceptable.

This has been a research topic of interest in recent years (see [1-3], and references therein). Most existing results however focus on the signed magnitude (SM) rounding and truncation schemes. Recently, some work on the two's complement (TC) truncation scheme has also appeared [5-7].

In what follows, an algorithm that may be used to check for limit cycles of a given digital filter implemented in fixed-point (FXP) arithmetic is proposed. It possesses a wide scope of applicability: The filter may be of any order; the quantization scheme may be arbitrary (including truncation and rounding schemes corresponding to SM and TC); and the accumulator may be of single- or double-length. For a given digital filter, bounds on the amplitude and period of possible limit cycles are developed. The algorithm is based on an exhaustive search over all these possibilities. Extending the same procedure to the entire linear stability region, one may now utilize it to obtain regions in filter coefficient space where the filter is globally asymptotically stable (g.a.s.). The robustness of the algorithm in terms of the absence of limit cycles with respect to filter coefficient perturbations may also be verified. The algorithm in [3], although developed with the same objectives in mind, is only applicable to filters implemented in direct form. In contrast, the proposed algorithm is applicable for the more general state-space (s.s.) formulation. Of course, the direct form then follows as a special case.

K. Premaratne and E.C. Kulasekere are with the Dept. of Elect. & Comp. Eng., P.O. Box 248294, Univ. of Miami, Coral Gables, FL 33124. P.H. Bauer is with the Dept. of Elect. Eng., Univ. of Notre Dame, Notre Dame, IN 46556.
L.J. Leclerc is with Ericsson Communications, Inc., Ville Mont-Royal, Québec, CANADA.
K.P. and P.H.B. gratefully acknowledge the support received from the Office of Naval Research (ONR) through grants N00014-94-1-0454 and N00014-94-1-0387, respectively.

II. Amplitude and period bounds on limit cycles.

In general, the quantization nonlinearity, $\mathcal{Q}[\cdot]$, satisfies

$$|x - \mathcal{Q}[x]| \leq \varrho \cdot q, \ \forall x \in \Re, \tag{1}$$

where ϱ is the normalized quantization error. In particular, for roundoff, $\varrho = 0.5$; for truncation, $\varrho = 1$. Note that, all filter parameters may be expressed as integer multiples of the quantization step size q. Hence, for convenience, we normalize q to unity. The quantization nonlinearity thus becomes an integer valued function, viz., $\mathcal{Q} : \Re \rightarrow \mathcal{Z}$. $\Re, \mathcal{Z}$, set of reals and integers. Typically, for all quantization schemes of interest, $\mathcal{Q}[0] = 0$. Consider a digital filter of order m in its minimal s.s. representation $\{A, B, C, D\}$:

$$\mathbf{x}(k+1) = A \cdot \mathbf{x}(k) + B \cdot \mathbf{u}(k); \tag{2}$$

$$\mathbf{y}(k) = C \cdot \mathbf{x}(k) + D \cdot \mathbf{u}(k), \tag{3}$$

where $\mathbf{x} \in \Re$ is the state, $\mathbf{u}$ is the input, and $\mathbf{y}$ is the output. Also, $A \in \Re^{m \times m}$. For addressing limit cycle performance, consider the zero input recursive state equation

$$\mathbf{x}(k+1) = A \cdot \mathbf{x}(k). \tag{4}$$

We only consider linearly stable filters, that is, all eigenvalues of A are inside the unit circle in $\mathcal{C}$ (set of complex numbers).

Now, under finite wordlength conditions, the pertinent quantization nonlinearity (4) may be modeled as

$$\mathbf{x}(k+1) = \mathcal{Q}[A \cdot \mathbf{x}(k)]. \tag{5}$$

Depending on whether the result of a product is stored with full precision or whether quantization is performed immediately after each product is computed, determines the effect of this nonlinearity. Considering (4), we get the following:
If the products can be stored with full precision, that is, if a double-length accumulator is available,

$$\mathbf{x}(k+1) = \begin{pmatrix} \mathcal{Q}[\sum_{j=1}^{m} a_{1j} \cdot x_j(k)] \\ \vdots \\ \mathcal{Q}[\sum_{j=1}^{m} a_{mj} \cdot x_j(k)] \end{pmatrix}. \tag{6}$$

Where $x_j(k)$ is the jth component of $\mathbf{x}(k)$. If the product is quantized immediately after each product is performed, that is, if only a single-length accumulator is available,

$$\mathbf{x}(k+1) = \begin{pmatrix} \mathcal{Q}[a_{11} \cdot x_1(k)] + \mathcal{Q}[a_{12} \cdot x_2(k)] + \ldots + \mathcal{Q}[a_{1m} \cdot x_m(k)] \\ \vdots \\ \mathcal{Q}[a_{m1} \cdot x_1(k)] + \mathcal{Q}[a_{m2} \cdot x_2(k)] + \ldots + \mathcal{Q}[a_{mm} \cdot x_m(k)] \end{pmatrix} \tag{7}$$

Noting (1), (6) and (7) may be expressed in a unified manner as

$$\mathbf{x}(k+1) = A \cdot \mathbf{x}(k) + \mathbf{e}(k), \quad \text{with} \quad |e_i(k)| \leq N \cdot \varrho, \tag{8}$$

where $\mathbf{e}(k) = \{e_i(k)\} \in \Re^m$ is the quantization error vector. If (6) is applicable, $N = 1$; if (7) is applicable, $N = m$.

We note that, (8) is a description of a *linear* system driven by the bounded input signal $\mathbf{e}(k)$. Hence, we have in fact converted the nonlinear systems in (6) and (7) into the linear system in (8).

Now, the transfer function between $\mathbf{e}(k)$ and $\mathbf{x}(k)$ is

$$\frac{\mathbf{X}(z)}{\mathbf{E}(z)} = (z \cdot I - A)^{-1} \in \Re(z)_{m \times m}, \tag{9}$$

where $\mathbf{X}$ and $\mathbf{E}$ are the z-transforms of $\mathbf{x}$ and $\mathbf{e}$, respectively and $\Re(z)_{m \times m}$ the set of matrices of size $m \times m$ over the rational polynomials in $z \in \mathcal{C}$. $\mathbf{I}$ is the identity matrix. This, when expanded, becomes

$$\frac{\mathbf{X}(z)}{\mathbf{E}(z)} = \begin{pmatrix} H_{11}(z) & H_{12}(z) & \dots & H_{1m}(z) \\ \vdots & \vdots & \ddots & \vdots \\ H_{m1}(z) & H_{m2}(z) & \dots & H_{mm}(z) \end{pmatrix}$$

where $H_{ij}(z) \in \Re(z)$. Hence,

$$X_i(z) = \sum_{j=1}^{m} H_{ij}(z) \cdot E_j(z), \; i = 1, 2, \dots, m, \tag{10}$$

where $\mathbf{X}(z) = \{X_i\}$ and $\mathbf{E}(z) = \{E_j\}$.
Therefore $x_i(k) = \sum_{j=1}^{m} h_{ij}(k) * e_j(k)$, $i = 1, 2, \dots, m$, where $h_{ij}(k)$ is the impulse response of $H_{ij}(z)$. Hence

$$x_i(k) = \sum_{j=1}^{m} \sum_{r=0}^{\infty} h_{ij}(r) \cdot e_j(k-r), \; i = 1, 2, \dots, m. \tag{11}$$

Noting $|e_j(k)| \leq N \cdot \varrho$, from (11), we get

$$|x_i(k)| \leq N \cdot \varrho \cdot \sum_{j=1}^{m} \sum_{k=0}^{\infty} |h_{ij}(k)|. \tag{12}$$

Therefore

$$M_i = N \cdot \varrho \cdot \sum_{j=1}^{m} \sum_{k=0}^{\infty} |h_{ij}(k)|, \; i = 1, 2, \dots, m. \tag{13}$$

M_i is the upper bound for the absolute value of $x_i(k)$. To estimate a useful upper bound for x_i, we need to compute $\sum_{j=1}^{m} \sum_{k=0}^{\infty} |h_{ij}(k)|$ for a given filter. We address this now.

Consider the transfer function $H_{ij}(z)$.

All poles of $H_{ij}(z)$ are distinct: $H_{ij}(z)$ may be expressed as

$$H_{ij}(z) = K_{ij} + \frac{r_{ij}^{(1)}}{1 - P_1^{(1)} z^{-1}} + \dots + \frac{r_{ij}^{(m)}}{1 - P_m^{(m)} z^{-1}},$$

where $r_{ij}^{(p)}, P_\ell^{(q)} \in \mathcal{C}$ and $K_{ij} \in \Re$, $r_{ij}^{(k)}$ is the kth residue of $H_{ij}(k)$ and P_i the poles. Taking inverse z-transform, we get

$$h_{ij}(k) = K_{ij} \cdot \delta(k) + r_{ij}^{(1)} [P_1^{(1)}]^k + \dots + r_{ij}^{(m)} [P_m^{(m)}]^k,$$

where $\delta(k)$ is the Dirac delta function. Therefore

$$\sum_{k=0}^{\infty} |h_{ij}(k)| \leq |K_{ij}| + |r_{ij}^{(1)}|(1 - |P_1^{(1)}|)^{-1} + \dots + |r_{ij}^{(m)}|(1 - |P_m^{(m)}|)^{-1}$$

This, when expanded, gives

$$\sum_{j=1}^{m} \sum_{k=0}^{\infty} |h_{ij}(k)| \leq \sum_{j=1}^{m} |K_{ij}| + \dots + (1 - |P_m^{(m)}|)^{-1} \cdot \sum_{j=1}^{m} |r_{ij}^{(m)}|.$$

for $i = 1, 2, \dots, m$. Hence

$$|x_i(k)| \leq N \cdot \varrho \cdot \left\{ \sum_{j=1}^{m} |K_{ij}| + \dots + (1 - |P_m^{(m)}|)^{-1} \cdot \sum_{j=1}^{m} |r_{ij}^{(m)}| \right\}. \tag{14}$$

for $i = 1, 2, \dots, m$. Convergence of this follows from linear stability.

Remark. The method in [3] tends to be easier to implement and more general with regards to its capability of handling poles of higher multiplicity. However, in our experience, the technique described above often leads to lower upper bounds. Note that, the technique in [3] utilizes an interpretation that involves a cascade of first-order sections whereas the technique above utilizes a parallel combination. Of course, no *one* technique will provide a lower bound for *all* situations. If computer cost is of concern, one can run both techniques and utilize the lower value of the bound.

$H_{ij}(z)$ contains a pole with multiplicity γ: Let this pole of multiplicity γ be P. Then,

$$H_{ij}(z) = K_{ij} + \frac{r_{ij}^{(1)}}{(1 - Pz^{-1})} + \frac{r_{ij}^{(2)}}{(1 - Pz^{-1})^2} + \dots + \frac{r_{ij}^{(\gamma)}}{(1 - Pz^{-1})^\gamma},$$

This analysis differs for the general term $r_{ij}^{(\zeta)}/(1 - Pz^{-1})^\zeta$ where $\zeta = 2, 3, \dots, \gamma$. At this point, due mainly to its ease of implementation, we utilize the technique in [3], by considering the general term and taking the inverse z-transform,

$$\frac{r_{ij}^{(\zeta)}}{(1 - Pz^{-1})(1 - Pz^{-1}) \dots (1 - Pz^{-1})} = r_{ij}^{(\zeta)} \cdot \left[\frac{1}{1 - |P|} \right]^\zeta$$

This expression is now substituted for the pole of multiplicity γ.

Lemma 1: The zero input response of the state $\mathbf{x}(k)$ of the digital filter described by eqn (6) or (7) is periodic. Its period T satisfies

$$T \leq \prod_{i=1} (2 \cdot \hat{M}_i + 1) = T_{max}, \tag{15}$$

where $\hat{M}_i$ is the largest integer not more than M_i given by eqn (13)

Proof: Consider eqn (6) or (7). The steady-state solution of each state $x_i(k)$ will satisfy $|x_i(k)| \leq M_i$, $\forall k, i = 1, 2 \dots, m$. Under FXP arithmetic, $\mathbf{x}(k) \in \mathcal{Z}$, and hence, $|x_i(k)| \leq \hat{M}_i$. $x_i(k)$ can therefore take only a finite number of values, namely, $(2 \cdot \hat{M}_i + 1)$. Hence, $\mathbf{x}(k)$ can take only

a finite number of values, namely, $\prod_{i=1}(2 \cdot \hat{M}_i + 1)$. Note that, the current state vector $\mathbf{x}(k)$ uniquely determines the next state vector $\mathbf{x}(k+1)$ through the function $\mathcal{Q}[\cdot]$. Thus, $\mathbf{x}(k)$ must be periodic in k. Its period is bounded by

$$T_{max} = \prod_{i=1}(2 \cdot \hat{M}_i + 1). \tag{16}$$

III. Algorithm Description

We now formulate the theoretical basis for the algorithm and discuss some of its computational aspects.

Definition 1: The digital filter realization in (8) is said to be globally asymptotically stable (g.a.s.) if and only if, for any initial state $\mathbf{x}(0) \in \mathcal{Z}$ with $\|\mathbf{x}(0)\|_\infty \leq B$, where $B \in \mathcal{Z}_+$, there exists $L \in \mathcal{Z}_+$ such that $\mathbf{x}(k) = \mathbf{0}$, where $\mathbf{0}$ is the null matrix, for $k \geq L$.

Remark. Typically, g.a.s. is taken to hold when $\mathbf{x}(k) \to \mathbf{0}$ as $k \to \infty$ (under the conditions above). However, due to the finite wordlength available, the filter behaves as a finite state machine, and Definition 1 suffices.

Lemma 2: Consider $\eta > 0$ and any initial state vector $\mathbf{x}(0)$ such that $|x_i(0)| \leq B_i$ with $B_i > \hat{M}_i$, for $i = 1, 2, \ldots, m$. Then, there exists a sufficiently large positive number $\mathcal{L}$ such that the digital filter in (6) or (7) satisfies $|x_i(k)| \leq \hat{M}_i + \eta$, $\forall k \geq \mathcal{L}$, for $i = 1, 2, \ldots, m$.

Proof: Since A is assumed to be linearly stable, the digital filter in (8) is in fact g.a.s. Hence, (8) will yield a set of nonhomogeneous linear shift-invariant difference equations which will have its solution in two parts: A steady-state solution $\mathbf{s}(k)$ and a transient solution $\mathbf{t}(k)$. Clearly, with g.a.s., given $\eta > 0$, we can choose k sufficiently large, say, $k \geq \mathcal{L}$, such that $\max|t_i(k)| < \eta$, for $i = 1, 2, \ldots, m$. Since $\hat{M}_i \in \mathcal{Z}_+$, for $k \geq \mathcal{L}$, $\hat{M}_i + \eta$ will act as a true upper bound for $x_i(k)$ in (8). □

Hence, it suffices to check the state vectors in

$$\mathcal{S}^{(0)} = \left\{\mathbf{x}(k) \in \mathcal{Z} \,\middle|\, |x_i(k)| \leq \hat{M}_i,\ i = 1, 2, \ldots, m\right\}, \tag{17}$$

to see if they are mapped to $\mathbf{0}$ by (8) after a finite number of iterations.

Computational Aspects : The computations within the algorithm are carried out in two stages. Initially, all vectors $\mathbf{x}(k) \in \mathcal{S}^{(0)}$ which map to $\mathbf{0}$ in less than T_{max} recursions—(after all, if limit cycles exist, the maximum period is T_{max})—are eliminated from $\mathcal{S}^{(0)}$. The remaining vectors in $\mathcal{S}^{(0)}$ are then further checked for convergence (see Section B).

Section A. Consider the set $\mathcal{V}^{(1)}$, where

$$\mathcal{V}^{(1)} = \left\{\mathbf{x}(k) \in \mathcal{S}^{(0)} | \mathcal{Q}[A \cdot \mathbf{x}(k)] = \mathbf{0}\right\}, \tag{18}$$

Hence, $\mathcal{V}^{(1)}$ consists of all the vectors $\mathbf{x}(k) \in \mathcal{S}^{(0)}$ that map to $\mathbf{0}$ in one and only one iteration of (6) or (7). Any other convergent vector in $\mathcal{S}^{(0)}$ must map to $\mathcal{V}^{(1)}$ prior to reaching $\mathbf{0}$. Hence, form

$$\mathcal{S}^{(1)} = \mathcal{S}^{(0)} \setminus \mathcal{V}^{(1)}. \tag{19}$$

Note that, $\mathcal{K}[\mathcal{S}^{(1)}] = \mathcal{K}[\mathcal{S}^{(0)}] - \mathcal{K}[\mathcal{V}^{(1)}]$. In fact, $\mathcal{K}[\mathcal{S}^{(0)}] = T_{max}$. $\mathcal{K}[\cdot]$ defines the cardinality of a set.

Furthermore, any vector in $\mathcal{S}^{(1)}$ which is mapped to $\mathcal{V}^{(1)}$ by (6) or (7) in one iteration will also converge to $\mathbf{0}$. Hence, form

$$\mathcal{V}^{(2)} = \left\{\mathbf{x}(k) \in \mathcal{S}^{(1)} | \mathcal{Q}[A \cdot \mathbf{x}(k)] \in \mathcal{V}^{(1)}\right\}. \tag{20}$$

Hence, $\mathcal{V}^{(2)}$ consists of all the vectors $\mathbf{x}(k) \in \mathcal{S}^{(1)}$ that map to $\mathbf{0}$ in exactly two iterations of (6) or (7). Hence, form

$$\mathcal{S}^{(2)} = \mathcal{S}^{(1)} \setminus \mathcal{V}^{(2)}. \tag{21}$$

Note that, $\mathcal{K}[\mathcal{S}^{(2)}] = \mathcal{K}[\mathcal{S}^{(0)}] - \mathcal{K}[\mathcal{V}^{(1)}] - \mathcal{K}[\mathcal{V}^{(2)}]$.
Likewise, we get the following sets: For $L = 1, 2, \ldots, T_{max}$,

$$\mathcal{V}^{(L)} = \left\{\mathbf{x}(k) \in \mathcal{S}^{(L-1)} | \mathcal{Q}[A \cdot \mathbf{x}(k)] \in \mathcal{V}^{(L-1)}\right\}, \tag{22}$$

and

$$\mathcal{S}^{(L)} = \mathcal{S}^{(L-1)} \setminus \mathcal{V}^{(L)}. \tag{23}$$

Note that, $\mathcal{K}[\mathcal{S}^{(L)}] = \mathcal{K}[\mathcal{S}^{(0)}] - \sum_{i=1}^{L} \mathcal{K}[\mathcal{V}^{(i)}]$.

The conditions under which this construction is terminated and their implications are as follows:
(1) If

$$\mathcal{K}[\mathcal{S}^{(L)}] = \emptyset, \quad \text{for some} \quad L = 1, 2, \ldots, T_{max} - 1, \tag{24}$$

all vectors in $\mathcal{S}^{(0)}$ are convergent.
(2) If

$$\mathcal{K}[\mathcal{V}^{(L)}] = \emptyset, \quad \text{for some} \quad L = 1, 2, \ldots T_{max}, \tag{25}$$

then

$$\mathcal{S}^{(i)} = \mathcal{S}^{(L-1)}, \quad \text{for} \quad i = L, L+1, \ldots, T_{\max}. \tag{26}$$

Under this situation, the remaining vectors in $\mathcal{S}^{(L-1)}$–there are $\mathcal{K}[\mathcal{S}^{(L-1)}]$ of them–will be further checked for convergence (see Section B).

Remark. Upon a little reflection, one notices that $\mathcal{V}^{(T_{max})}$ must either be empty or contain one and only one vector from $\mathcal{S}^{(0)}$.

Section B. Although the reverse mapping procedure outline above reduces the computational complexity considerably, it may not capture all the vectors in $\mathcal{V}^{(L)}$, $L = 1, 2, \ldots, T_{max}$, that map to $\mathbf{0}$ within T_{max} iterations. This is due to the fact that, there may be vectors in $\mathcal{V}^{(L)}$ that map to $\mathbf{0}$ through a vector not belonging to $\mathcal{S}^{(0)}$! Hence, when encountered with condition (2) above, convergence of each remaining vector in $\mathcal{S}^{(L-1)}$ is determined by checking whether it is mapped to $\mathbf{0}$ in less than T_{max} through either (6) or (7), whichever is applicable. This exhaustive technique is in fact an extension of that given in [3] to digital filters represented in their s.s. realization. However, we must emphasize the significant computational advantage gained by first invoking the mapping procedure in Section A. Assuming condition (2) has occurred, let

$$\mathcal{S}^{(L)} = \left\{\mathbf{x}_i^{(L)},\ i = 1, 2, \ldots, \mathcal{K}[\mathcal{S}^{(L)}]\right\}. \tag{27}$$

Note that, when condition (2) has occurred, from (26), $\mathcal{S}^{(L-1)} = \mathcal{S}^{(L)}$. For each vector $\mathbf{x}_i^{(L)} \in \mathcal{S}^{(L)}$, construct the orbit $\mathcal{O}_i^{(L)}$ consisting of all state vectors $\mathbf{x}_i^{(L)}(j)$, for $j = 1, 2, \ldots, T_{max}$, that are consecutively generated by (6) or (7) (whichever is applicable) with $\mathbf{x}_i^{(L)}$ as the initial state, that is, $\mathbf{x}_i^{(L)} = \mathbf{x}_i^{(L)}(0)$. For each $i = 1, 2, \ldots, \mathcal{K}[\mathcal{S}^{(L)}]$, the conditions under which the construction of each orbit $\mathcal{O}_i^{(L)}$ is terminated and their implications are as follows:
(1) If

$$\mathbf{x}_i^{(L)}(j) = 0, \quad \text{for some} \quad j = 1, 2, \ldots, T_{max}, \tag{28}$$

then, $\mathbf{x}_i^{(L)}$ together with each vector in the orbit $\mathcal{O}_i^{(L)}$ will be convergent.
(2) If

$$\mathbf{x}_i^{(L)}(j) = \mathbf{x}_i^{(L)}(k), \quad \text{for} \quad j \neq k, \tag{29}$$

then $\mathbf{x}_i^{(L)}$ gives rise to limit cycles.
Remark. These are in fact the only conditions that can occur when either (6) or (7) generate the orbit.
Note : An analysis was carried out to determine the robustness regions under single- and double- length environments the results of which are found in [8].

IV. Some Examples

The proposed algorithm is applied to a dense grid in the coefficient space to obtain the total g.a.s. region. With the robustness region due to the variation of the coefficients, each point in the coefficient space is associated with a neighborhood where the filter is stable. A 10-bit wordlength is assumed for all computations. Therefore, the filter coefficients are quantized to a multiple of 2^{-10}. Within the linear stability region, dark areas indicate filters that possess limit cycles.

The results provided correspond to the most commonly encountered quantization schemes, namely, SM roundoff, SM truncation, and TC truncation schemes. In all cases, both single- and double-length accumulator implementation results were analyzed. All results are provided for a second-order filter. All results given in [3] for direct form filters were also verified using this algorithm.

Results for minimum norm realization of digital filters
Stability of digital filters in its minimum norm realization was also investigated. The coefficient matrix, in such a case, is $A = \begin{bmatrix} \sigma & \omega \\ -\omega & \sigma \end{bmatrix}$.

The results for the SM roundoff for both single- and double-length accumulator environments were verified [9]. The stable region for SM truncation scheme for both single- and double-length accumulator environments span the entire linear stability region $\sigma^2 + \omega^2 < 1$.

For TC truncation, with double-length accumulator, the g.a.s. region is in Figure (1a). This in fact improves on the previously known results in [7]. For instance, the series of points that satisfy $\sigma < 0$ and $\omega = \pm\sigma$., are also limit cycle free. The following coefficient matrix belongs to this class:

$$A = \begin{bmatrix} -\frac{672}{1024} & \frac{672}{1024} \\ -\frac{672}{1024} & -\frac{672}{1024} \end{bmatrix}. \tag{30}$$

To the authors' knowledge, no previous results are available for TC truncation in a single-length accumulator environment. The region of g.a.s. is in Figure (1b).

VII. Conclusion

A new algorithm capable of determining g.a.s. of any FXP digital filter in its s.s. formulation has been presented. The algorithm is applicable independent of the nonlinearity, number of nonlinearities, and order of filter. In most cases, the proposed algorithm is found to provide tighter bounds on the amplitude of limit cycles. Significant improvements over existing results for the TC truncation schemes in both single- and double-length accumulator environments have been presented. Current research is directed towards establishing regions within which limit cycles of a pre-specified period or bound exists.

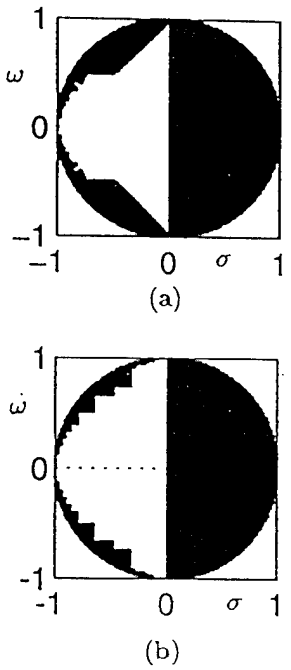


Figure 1: The region where a filter with twos complement truncation is free of limit cycles. (a) Double (b) Single ,length accumulator

References

[1] T. Claasen, W.F.G. Mecklenbräuker, and J.B.H. Peek, "Frequency domain criteria for the absence of zero-input limit cycles in nonlinear discrete-time systems with applications to digital filters," *IEEE Trans. Circ. Syst.*, vol. CAS-22, no. 3, pp. 232-239, Mar. 1974.

[2] K.T. Erickson and A.N. Michel, "Stability analysis of fixed-point digital filters using computer generated Lyapunov functions Part I: Direct form and coupled form filters," *IEEE Trans. Circ. Syst.*, vol. CAS-32, no. 2, pp. 113-131, Feb. 1985.

[3] P.H. Bauer and L.J. Leclerc, "A computer-aided test for the absence of limit cycles in fixed-point digital filters," *IEEE Trans. Sig. Proc.*, vol. 39. no. 11, pp. 2400-2409, Nov. 1991.

[4] J.F. Kaiser, "Some special practical considerations in the realization of linear digital filters,"
Proc. 3rd Allerton Ann. Conf., Circ. Syst. Theory, pp. 100-104, 1965.

[5] T. Bose and D.P. Brown, "Limit cycles in zero input digital filters due to two's complement quantization," *IEEE Trans. Circ. Syst.*, vol. CAS-37, no. 4, pp. 568-571, Month April 1990.

[6] A. Lepschy, G.A. Mian and U. Viaro, "Effects of quantization in second order fixed point digital filters with two's complement truncation quantization," *IEEE*
Trans. Circ. Syst., vol. CAS-35, no. 4, pp. 461-466, April 1988.

[7] T. Bose, Mei-Qin Chen and Frank Brammer "Stability of normal form digital filters with Two's complement quantization," *ICASSP '91 Proc.*, , pp. 1889-1892.

[8] K. Premeratne and E.C. Kulasekere "An exhaustive search algorithm for checking limit cycle behavior of digital filters," *IEEE Trans. Sig. Proc.*, In preparation.

[9] Trân-Thông and B. Liu, "A contribution to the stability analysis of second order direct form digital filters with magnitude truncation," *IEEE Trans. Acoust., Speech, Sig. Proc.*, vol. ASSP-35, no. 8, pp. 1207-1210, Aug. 1987.

Proceedings of the 38th Midwest Symposium on Circuits and Systems (MWSCAS'95), Rio de Janeiro, Brazil, August 1995, to appear.

ZERO-CONVERGENCE OF 2-D ROESSER STATE SPACE MODELS IMPLEMENTED IN FLOATING POINT ARITHMETIC

Peter H. Bauer
Department of Electrical Engineering
University of Notre Dame
Notre Dame, IN 46556
Tel: (219) 631-8015
e-mail: pbauer@mars.ee.nd.edu

Kamal Premaratne
Department of Electrical & Computer Engineering
University of Miami
Coral Gables, FL 33124
Tel: (305) 284-4051
e-mail: kprema@umiami.ir.miami.edu

Abstract:

Zero input asymptotic response behavior of general order 2-D digital filters with floating point arithmetic is investigated. In particular, conditions for the absence of so-called R1 and R2 responses (large amplitude limit cycles) are provided for 2-D first quarter-plane causal filters.

I. Introduction

Recently, floating point arithmetic has become popular for a number of digital signal processing applications. The implementation of digital filters in floating point format is especially attractive due to the high dynamical range and the high-level programming tools available.

Previous work on the convergence behavior of floating point digital filters concentrated on 1-D second order system [1]. Some results on direct form filters are also available [2]. However, to the authors' knowledge, the case of general order 1-D or 2-D state space models has not been tackled.

This paper provides such an analysis which can be applied to any digital filter structure of arbitrary order and dimension one or two. In order to avoid distinguishing among a number of reformatting and quantization schemes, the result introduced in this paper takes a parameterization approach to the error description. This allows to apply the derived result to any type of floating point format.

II. Preliminaries

Consider the Roesser model for the first quarter-plane causal 2-D system:

$$\begin{pmatrix} \underline{\hat{x}}^h(n_1+1,n_2) \\ \underline{\hat{x}}^v(n_1,n_2+1) \end{pmatrix} = \begin{pmatrix} A_{hh} & A_{hv} \\ A_{vh} & A_{vv} \end{pmatrix} \begin{pmatrix} \underline{\hat{x}}^h(n_1,n_2) \\ \underline{\hat{x}}^v(n_1,n_2) \end{pmatrix} \tag{1}$$

$$\begin{aligned} A &= \begin{pmatrix} A_{hh} & A_{hv} \\ A_{vh} & A_{vv} \end{pmatrix}, \; A \in \Re^{(N_1+N_2)\times(N_1+N_2)} \\ & A_{hh} \in \Re^{N_1 \times N_1} \\ & A_{vv} \in \Re^{N_2 \times N_2} \end{aligned} \tag{2}$$

The submatrices A_{vh} and A_{hv} are of the appropriate dimensions. The vectors $\underline{\hat{x}}^h$ and $\underline{\hat{x}}^v$ are horizontally and vertically propagating state vectors of the ideal system, respectively.

For floating point realizations of (1), the following error model describes the system behavior:

$$\begin{pmatrix} \underline{x}^h(n_1+1,n_2) \\ \underline{x}^v(n_1,n_2+1) \end{pmatrix} = \begin{pmatrix} A_{hh} & A_{hv} \\ A_{vh} & A_{vv} \end{pmatrix} \begin{pmatrix} \underline{x}^h(n_1,n_2) \\ \underline{x}^v(n_1,n_2) \end{pmatrix} + \begin{pmatrix} \underline{e}^h(n_1,n_2) \\ \underline{e}^v(n_1,n_2) \end{pmatrix} \tag{3}$$

where $\underline{e}^h(n_1,n_2) \in \Re^{N_1}$, $\underline{e}^v(n_1,n_2) \in \Re^{N_2}$ are the error vectors for the horizontally and vertically propagating states, respectively.

We also need to define the following transfer matrices:

$$\begin{pmatrix} \underline{X}^h(z_1,z_2) \\ \underline{X}^v(z_1,z_2) \end{pmatrix} = \begin{pmatrix} H^{hh}(z_1,z_2) & H^{hv}(z_1,z_2) \\ H^{vh}(z_1,z_2) & H^{vv}(z_1,z_2) \end{pmatrix} \begin{pmatrix} \underline{E}^h(z_1,z_2) \\ \underline{E}^v(z_1,z_2) \end{pmatrix} \tag{4}$$

where

$$\begin{pmatrix} H^{hh}(z_1,z_2) & H^{hv}(z_1,z_2) \\ H^{vh}(z_1,z_2) & H^{vv}(z_1,z_2) \end{pmatrix} = \left(\begin{bmatrix} z_1 I_1 & \phi \\ \phi & z_2 I_2 \end{bmatrix} - A \right)^{-1} \tag{5}$$

In Equation (4), $\underline{X}^h(z_1,z_2)$ and $\underline{X}^v(z_1,z_2)$ are the z-transforms of the states $\underline{x}^h(n_1,n_2)$ and $\underline{x}^v(n_1,n_2)$, respectively. The transforms $H^{hh}(z_1,z_2), H^{hv}(z_1,z_2), H^{vh}(z_1,z_2)$ and $H^{vv}(z_1,z_2)$ are transfer submatrices of dimensions $N_1 \times N_1, N_1 \times N_2, N_2 \times N_1, N_2 \times N_2$, respectively. $\underline{E}^h(z_1,z_2)$ and $\underline{E}^v(z_1,z_2)$ are the 2-D z-transforms of the error signal vectors $\underline{e}^h(n_1,n_2)$ and $\underline{e}^v(n_1,n_2)$, respectively. Furthermore, in (5), I_1 and I_2 denote identity matrices of dimensions $N_1 \times N_1$ and $N_2 \times N_2$, respectively.

The components $H^{hh}(z_1,z_2), H^{hv}(z_1,z_2), H^{vh}(z_1,z_2)$ and $H^{vv}(z_1,z_2)$ are 2-D transforms denoted by $H_{ij}^{hh}(z_1,z_2), H_{ij}^{hv}(z_1,z_2), H_{ij}^{vh}(z_1,z_2)$ and $H_{ij}^{vv}(z_1,z_2)$, respectively.
Denoting $\mathcal{Z}\{\cdot\}$ as the 2-D z-transform, we define the following impulse responses:

$$H_{ij}^{hh}(z_1,z_2) = \mathcal{Z}\{h_{ij}^{hh}(n_1,n_2)\};\ i=1,\cdots,N_1;\quad j=1,\cdots,N_1. \tag{6}$$

$$H_{ij}^{hv}(z_1,z_2) = \mathcal{Z}\{h_{ij}^{hv}(n_1,n_2)\};\ i=1,\cdots,N_1;\quad j=1,\cdots,N_2. \tag{7}$$

$$H_{ij}^{vh}(z_1,z_2) = \mathcal{Z}\{h_{ij}^{vh}(n_1,n_2)\};\ i=1,\cdots,N_2;\quad j=1,\cdots,N_1. \tag{8}$$

$$H_{ij}^{vv}(z_1,z_2) = \mathcal{Z}\{h_{ij}^{vv}(n_1,n_2)\};\ i=1,\cdots,N_2;\quad j=1,\cdots,N_2. \tag{9}$$

Next, we define the l_1-measures for each component of the transfer function submatrices:

$$\tilde{H}_{ij}^{vv} = \sum_{n_1=0}^{\infty}\sum_{n_2=0}^{\infty} |h_{ij}^{vv}(n_1,n_2)|;\ i,j=1,\cdots,N_1. \tag{10}$$

$$\tilde{H}_{ij}^{hv} = \sum_{n_1=0}^{\infty}\sum_{n_2=0}^{\infty} |h_{ij}^{hv}(n_1,n_2)|;\ i=1,\cdots,N_1;\quad j=1,\cdots,N_2. \tag{11}$$

$$\tilde{H}_{ij}^{vh} = \sum_{n_1=0}^{\infty}\sum_{n_2=0}^{\infty} |h_{ij}^{vh}(n_1,n_2)|;\ i=1,\cdots,N_2;\quad j=1,\cdots,N_1. \tag{12}$$

$$\tilde{H}_{ij}^{vv} = \sum_{n_1=0}^{\infty}\sum_{n_2=0}^{\infty} |h_{ij}^{vv}(n_1,n_2)|;\ i,j=1,\cdots,N_2. \tag{13}$$

Also:

$$\tilde{H}_i^h = \sum_{\nu=1}^{N_1} \tilde{H}_{i\nu}^{hh} + \sum_{\nu=1}^{N_2} \tilde{H}_{i\nu}^{hv} \tag{14}$$

$$\tilde{H}_j^v = \sum_{\nu=1}^{N_1} \tilde{H}_{j\nu}^{vh} + \sum_{\nu=1}^{N_2} \tilde{H}_{j\nu}^{vv} \tag{15}$$

From [1,2] it is known that the following four state response types are encountered in floating point digital filters under zero input, if the linear filter is stable:

R1: an unbounded state response, eventually leading to overflow conditions.
R2: a bounded state response
R3: a bounded state response in underflow
R4: a zero-convergent response

III. The Main Result

The following theorem can now be formulated:

Theorem: A floating point implementation of the system in (1) for any finitely extended input signal and/or non-zero finitely extended initial conditions will produce a response type R3, if the mantissa length l_m satisfies

$$l_m \geq 2 + \log_2 \tilde{H} + \log_2 C \tag{16}$$

where $\tilde{H} = \max_{i,j}(\tilde{H}_i^h, \tilde{H}_j^v)$ and C is an implementation dependent constant.

Proof: The proof is rather lengthy and will be supplied in the final version of the paper.

Formally, this result is similar to previous results on direct forms [1] and second order state-space systems [2]. In this case, the stability margin enters the inequality through $\tilde{H}$, which is a somewhat complicated measure of the degree of stability of the system. For an unstable system $\tilde{H} \to \infty$, and for any stable system we have $\tilde{H} < \infty$. The constant C relates the magnitude of the state-variables to the error bound. This number is usually small and is directly affected by the entries of the A-matrix and the floating point format.

IV. Conclusion

This paper presents a condition on the mantissa length of a 2-D floating point digital filter of arbitrary order, which ensures convergence of the state-response into underflow, independent of the initial conditions. The mantissa length is linked to the margin of stability of the linear system as measured by $\tilde{H}$. It is also dependent on the realization itself. It should be noted that the response types R2 and R3 in the 2-D (and m-D) case do not

need to be periodic [3].

References

[1] P. H. Bauer, "Absolute Response Error Bounds for Floating Point Digital Filters in State Space Representation", *IEEE International Symposium on Circuits and Systems*, Chicago, May 3-6, 1993, pp. 619-622.

[2] P. H. Bauer and J. Wang, "Limit Cycle Bounds for Floating Point Implementations of Second Order Recursive Digital Filters", *IEEE Trans. on Circuits and Systems - Part II: Analog and Digital Signal Processing*, Vol. 40, No. 8, pp. 493-501, Aug. 1993.

[3] P.H. Bauer and E. I. Jury, "Nonperiodic Modes in 2-D Recursive Digital Filters under Finite Wordlength Effects", *IEEE Trans. on Circ. and Syst.*, Vol. 36, No. 7, pp. 11032-1035, 1989.

IASTED International Conference on Modelling and Simulation, Colombo, Sri Lanka, in review, July 1995.

iasted*.tex

Digital Simulation of Nonlinear Systems Using Delta-Operator Based Numerical Schemes

KAMAL PREMARATNE, Department of Electrical and Computer Engineering, P.O. Box 248294, University of Miami, Coral Gables, FL 33124 USA,
Tel: +1(305)284 4051; Fax: +1(305)284 4044; email: kprema@umiami.ir.miami.edu, and
PETER H. BAUER, Laboratory for Signal and Image Analysis (LISA), Department of Electrical Engineering, University of Notre Dame, Notre Dame, IN 46556 USA,
Tel: +1(219)631 8015; Fax: +1(219)631 4393; email: pbauer@mars.ee.nd.edu.

Extended Abstract

This extended abstract is being submitted for possible presentation at the *IASTED International Conference on Modelling and Simulation,* Colombo, Sri Lanka, July 26-28, 1995.

Introduction

Traditional control and signal processing algorithms based on shift-operator (or, q-operator) are ill-conditioned in high performance applications that involve fast sampling/shorter wordlength [1]. In these situations, q-operator based discrete-time implementations (or, q-systems) are extremely sensitive to uncertainties inherent in modelling and parameter representation (in particular, with shorter wordlength).

Use of incremental difference operator or delta-operator (or, δ-operator) can provide an effective solution to such difficulties [1]. Compared to q-systems, δ-operator based implementations (or, δ-systems) can provide superior performance with respect to (a) coefficient sensitivity of frequency response [1], and (b) quantization noise propagation [2]. Due mainly to these, and also due to the possibility of a unified treatment of both continuous- and discrete-time systems, work on δ-systems has recently attracted considerable attention (see [1-5], and references therein).

Problem Statement

Since δ-operator can offer several important advantages over q-operator for linear, time-invariant one-dimensional (1-D) systems, would similar advantages hold true for more general classes of systems? Work on *linear,* multi-dimensional (m-D) systems indicate that this may indeed be the case [5]. In this paper, we investigate the applicability of δ-operator based numerical schemes for simulation of *nonlinear* systems.

Delta-Operator Based Numerical Scheme

q-Operator Based Numerical Scheme. We consider the computation of solution orbit of a nonlinear system of the type

$$q[\mathbf{x}](n) = \mathbf{f}_q(\mathbf{x}(n), \mathbf{a}_q), \tag{1}$$

where $q[\mathbf{x}](n) = \mathbf{x}(n+1)$. Here, $\mathbf{x}(n)$ is the state orbit $\mathbf{x} \in \Re^m$ at instant n and

Kamal Premaratne and Peter H. Bauer gratefully acknowledge the support provided by the Office of Naval Research (ONR) through the grants N00014-94-1-0454 and N00014-94-1-0387, respectively.

$\mathbf{a}_q = [a_{1_q}, \ldots, a_{M_q}]^T \in \Re^M$ refer to system parameters that are *actually stored* within the computer while performing the iteration.

δ-Operator Based Numerical Scheme. The proposed δ-operator based scheme of the same nonlinear system in (1) is

$$\begin{aligned} \delta[\mathbf{x}](n) &= \mathbf{f}_\delta(\mathbf{x}(n), \mathbf{a}_\delta) \quad \text{(Intermediate equation)} \\ q[\mathbf{x}](n) &= \mathbf{x}(n) + \Delta \cdot \delta[\mathbf{x}](n) \quad \text{(Update equation)}, \end{aligned} \tag{2}$$

where $\delta[\mathbf{x}](n) = (q[\mathbf{x}](n) - \mathbf{x}(n))/\Delta$ and $\mathbf{f}_\delta(\mathbf{x}(n), \mathbf{a}_\delta) = (\mathbf{f}_q(\mathbf{x}(n), \mathbf{a}_q) - \mathbf{x}(n))/\Delta$. Here, Δ is an arbitrary positive real parameter (usually the grid size) and $\mathbf{a}_\delta = [a_{\delta_1}, \ldots, a_{\delta_M}]^T \in \Re^M$ again refer to system parameters that are *actually stored* within the computer.

Now, which of the schemes (1) or (2) yield superior coefficient sensitivity of its orbit with respect to perturbation of $\mathbf{a}_q$ or $\mathbf{a}_\delta$, respectively? This consideration is crucial in high performance, real-time applications that may require fast sampling/shorter wordlength. Of course, with infinite wordlength, both (1) and (2) yield identical results. In our development, the nonlinearity is taken to belong to $\mathcal{C}^1$, that is, it possesses first partial derivatives. Small perturbations are assumed.

CONTRIBUTIONS

The contributions of this paper are the following:

1. Development of coefficient sensitivity measures M_{FXP} and M_{FLP} for fixed-point (FXP) and floating-point (FLP), repectively. These take into account that in FXP, coefficient perturbation is approximately independent of its nominal value, while in FLP, it is approximately proportional.
2. FXP: M_{FXP} for δ-system is Δ times M_{FXP} for q-system. Hence, δ-system is superior under small grid size.
3. FLP: M_{FLP} for δ-system is superior than M_{FLP} for q-system if $|a_{i_q} - 1| \leq |a_{i_q}|$, $\forall i = 1, \ldots, M$. Here, a_{i_q} indicates the 'linear' term in the i-th equation of $\mathbf{f}_q$. We show that, typical digital equivalents of continuous-time nonlinear systems obtained under fast sampling routinely satsify this condition.
4. Similar comments hold true for linear systems, piecewise $\mathcal{C}^1$ nonlinear systems, and piecewise linear systems.

REFERENCES

[1] Goodwin, G.C., Middleton, R.H., and Poor, H.V. (1992). High speed digital signal processing and control. *Proc. IEEE,* **40**, 240-259.

[2] Li, G., and Gevers, M. (1993). Roundoff noise minimization using delta operator realizations. *IEEE Trans. Sig. Proc.,* **41**, 629-637.

[3] Premaratne, K., and Bauer, P.H. (1994). Limit cycles and asymptotic stability of delta-operator systems in fixed-point arithmetic. *Proc. IEEE Symp. Circ. Syst. (ISCAS'94),* London, UK, **2**, 461-464.

[4] Premaratne, K., and Jury, E.I. (1994). Tabular method for determining root distribution of delta-operator formulated real polynomials. *IEEE Trans. Auto. Cont.,* **39**, 352-355.

[5] Premaratne, K., Suarez, J., Ekanayake, M.M., and Bauer, P.H. (1994). Two-dimensional delta-operator formulated discrete-time systems: State-space realization and its coefficient sensitivity properties. *Proc. Midwest Symp. Circ. Syst. (MWSCS'94),* Lafayette, LA; *IEEE Trans. Sig. Proc.* in review.

SEMIANNUAL PERFORMANCE REPORT
GRANT NO's: N00014-94-1-0387 and N00014-94-1-0454

Summary of Phase P1 Results

Phase P1 consists of two tasks:

[T1] Task T1: Analysis and design of finite wordlength implementations of linear, time-invariant δ-Systems.

[T2] Task T3: 2-D and m-D δ-system models.

Major part of task T1 was carried out at the University of Notre Dame by Dr. Peter H. Bauer while major part of task T3 was carried out at the University of Miami by Dr. Kamal Premaratne. The project being an extensive collaborative effort, during this research work, the two PI's have been in constant contact.

The following is a summary of the phase P1 results.

Task T1: Analysis and Design of Finite Wordlength Implementations of Linear, Time-Invariant δ-Systems

The conclusions drawn from the work conducted for task T1 may be summarized as follows:

1. The Fixed-Point Arithmetic Case: When limit cycle performance is crucial, the q-operator implementation is preferrable. The δ-operator implementation is superior with regard to coefficient sensitivity issues.

2. The Floating-Point Arithmetic Case: Generally, the δ-operator implementation outperforms its q-operator counterpart. In particular, in high-order and high-speed applications, the δ-operator implementation is the best choice.

Prior to a more detailed exposition, first we provide qualitative justification for the above conclusion. The state equations of a δ-operator system can be written as:

$$\begin{aligned} \delta[\mathbf{x}](n) &= A_\delta \mathbf{x}(n) + B_\delta \mathbf{u}(n); \\ q[\mathbf{x}](n) &= \mathbf{x}(n) + \Delta \cdot \delta[\mathbf{x}](n). \end{aligned} \tag{T1.1}$$

where $\mathbf{x}$ and $\mathbf{u}$ are the state and input vectors, respectively. Here, Δ denote a positive real constant (typically, the sampling time). The symbol $\delta[\cdot]$ denotes the δ-operator, that is,

$$\delta[\mathbf{x}](n) = \frac{q[\mathbf{x}](n) - \mathbf{x}(n)}{\Delta} = \frac{q-1}{\Delta}\mathbf{x}(n), \tag{T1.2}$$

and $q[\cdot]$ denotes the usual q-operator, that is,

$$q[\mathbf{x}](n) = \mathbf{x}(n+1). \tag{T1.3}$$

The corresponding formulation of (T1.1) in terms of the q-operator is

$$q[\mathbf{x}](n) = A_q\mathbf{x}(n) + B_q\mathbf{u}(n), \tag{T1.4}$$

where

$$A_q = I + \Delta \cdot A_\delta \Longleftrightarrow A_\delta = \frac{A_q - I}{\Delta} \quad \text{and} \quad B_q = \Delta \cdot B_\delta \Longleftrightarrow B_\delta = \frac{B_q}{\Delta}. \tag{T1.5}$$

Now, given $\mathbf{x}$ and $\mathbf{u}$, both representations compute $q[\mathbf{x}]$ with a certain accuracy. Consider the δ-operator formulation in (T1.1). Here we encounter two errors:

1. The first is due to the computation of $\delta[\mathbf{x}]$, that is, the first equation in (T1.1). We will refer to this equation as the *intermediate equation.*

2. The second is due to the eventual computation of $q[\mathbf{x}]$, that is, the second equation in (T1.1). We will refer to this equation as the *update equation.*

Let us assume that the total error in computing $q[\mathbf{x}]$ is mainly due to the intermediate equation in (T1.1) (rather than the update equation). Then, by choosing Δ sufficiently small, the total error in computing $q[\mathbf{x}]$ will be approximately the error created by the update equation which is small!. In this case, the δ-operator representation has better finite wordlength properties than its q-operator counterpart in (T1.4).

If, however, the errors accumulated in the intermediate and the update equations in (T1.1) are comparable, $q[\mathbf{x}]$ computed through the δ-operator representation will show approximately the same error as that computed through its q-operator counterpart assuming Δ is sufficiently small. If Δ is not sufficiently smaller than one, the δ-operator representation will actually perform worse than the q-operator representation!

If the error introduced in the update equation is larger than that in the intermediate equation, the δ-operator representation would consistently perform worse!! In reality, this case is very unlikely to occur.

Next, a more detailed exposition follows.

T1.1 The Fixed-Point Arithmetic Case

We now discuss some of the results regarding the fixed-point (FXP) case. Here, our results

in fact indicate that, in case limit cycle behavior is crucial, the δ-operator representation is NOT suitable with this arithmetic scheme [1]. Such a case may occur when nonlinear systems are implemented through FXP δ-operator based schemes.

Zero-input limit cycles. Independent of Δ, zero-input limit cycles cannot be avoided in FXP δ-implementations. This is easily explained as follows: If Δ is chosen very small, the contribution from the intermediate equation being small (since $\delta[\mathbf{x}]$ is being multiplied by Δ), during the update equation, $q[\mathbf{x}]$ can be quantized to $\mathbf{x}$ creating a DC limit cycle, that is, an incorrect equilibrium point different from zero results. We emphasize that, most of the desirable properties of δ-operator implementations are based on a small Δ. We may also show that, if Δ is chosen larger (this case is of course somewhat less important), DC limit cycles will still exist. Hence, δ-operator representations cannot be implemented limit cycle free in FXP format! This fact is independent of the particular realization of the system.

Deadband size. Since δ-systems cannot be implemented limit cycle free in FXP format, it is of interest to investigate te the size of such limit cycles since, in certain situations, such small limit cycle amplitudes can be tolerated. It can be shown that, the magnitude of Δ determines the magnitude of the limit cycle. The smaller the Δ, the larger will be the deadband and hence the limit cycle magnitude. An approximate relationship regarding this is

$$\Delta \times \text{size of deadband} = 1, \tag{T1.6}$$

where the size of deadband is measured in multiples of the quantization step size. Here, the deadband corresponds to that obtained by considering the quantization of $\Delta \cdot \delta[\mathbf{x}]$. Therefore, the usual choice of a small Δ creates a larger deadband!

The input driven case. Although the input driven case is not part of the originally proposed work, some interesting results have been obtained. For small values of Δ, there exists a bounded input signal that does not allow control of the state trajectory. In other words, given sufficiently small Δ, the state trajectory may not be influenced by such an input signal.

The influence of the realization. First, it was necessary to develop a suitable scheme to investigate the effect of realization on the presence or absence of limit cycles. In this direction, for the q-operator case, a computer-based exhaustive search algorithm that checks for limit cycles (DC and/or oscillatory) has been developed [5].

As discussed before, we have shown that, a stable linear time-invariant δ-system cannot be implemented limit cycle free in FXP. The size of the deadband however also depends on the particular realization, that is, the structure of A_δ. Given a system transfer function, there are forms which minimize this deadband size with respect to some appropriately chosen measure. For example, in order to minimize DC limit cycle amplitude, one may choose the normal form (in terms of A_δ) as a suitable candidate.

The influence of quantization nonlinearity and its deadzone. Since a larger deadzone implies larger DC limit cycle amplitudes, the use of quantizers with reduced, or even zero, deadzone was therefore proposed. In investigating first-order systems, by reducing the deadzone, it was found that, existence of DC limit cycles can indeed be reduced. Unfortunately, other oscillatory limit cycles will be created. This phenomenon is due to the increased gain exhibited towards small input signals by the quantizer.

Scaling. As discussed above, we have shown that, independent of either the form of A_δ or the magnitude of Δ, a FXP implemented δ-system cannot be free of zero-input limit cycles. Hence, scaling cannot be offered as a possible solution.

T1.2 The Floating-Point Arithmetic Case

The floating-point (FLP) implementation of δ-systems is currently under investigation. The results obtained so far are very encouraging, and indicate that, quantization errors due to FLP arithmetic have a much smaller effect on the system behavior than in the FXP case. In fact, preliminary results show that, for δ-systems of order three and higher, errors in computing $q[\mathbf{x}]$ can be made significantly smaller than for the corresponding q-systems. This is because, for a FLP implementation of such a system, errors created through the intermediate equation are larger than those created through the update equation. As previously mentioned, in this situation, δ-systems behave better than their q-operator counterparts!

Limit cycles. In FLP arithmetic, a linearly stable time invariant system, under zero-input conditions, may exhibit four types of responses: A diverging response, an oscillatory periodic response of arbitrary magnitude, an oscillatory periodic response in underflow, or an asymptotically stable response. Only the last two response types are acceptable in practice. It is well known that, the last response type is in fact a very stringent requirement and is often not required in practice. Results so far obtained show that, when the requirements for a response in underflow are compared, the δ-system requires less wordlength than its q-system counterpart! This advantage in fact grows with the order of the system!!

Once the system reaches underflow conditions, the δ-system again exhibits DC limit cycles. However, if the exponent register is chosen sufficiently large, the amplitude of these oscillations can be made extremely small and hence, for all practical purposes, this problem is solved.

Deadband size. If the condition on the mantissa length that guarantees convergence into underflow is satisfied, then the deadband size will be very small. Hence, it can be neglected for all practical purposes. This assumes a properly chosen exponent register length since the exponent register length determines the dynamic range of underflow.

The Influence of the Nonlinearity. Unlike the FXP case, the characteristic of the nonlinearity has only a minor effect on the system behavior, significant differences being present only in underflow conditions

The Underflow case. In underflow, the δ-system seems to behave worse than its q-operator counterpart. This is mainly due to the fact that, a FLP system in underflow essentially performs very similar to a FXP system. However, as mentioned above, if the dynamic range of underflow is chosen properly, the system behavior in underflow is of little practical interest.

Block Floating-Point Arithmetic. Even for the q-operator case, results regarding block FLP implementations are lacking. Hence, investigations regarding block FLP implementation of δ-systems is in its early stages. In order to obtain a comparison between the two types of implementations, current research is geared towards obtaining results applicable for the q-operator case.

T1.3 The Multi-Dimensional Case

The results on one-dimensional (1-D) δ-operator implementations in FXP arithmetic directly carry over to the multi-dimensional (m-D) case. The existence of non-converging responses along the boundary of the causality region can easily be proven using the same type of argument used in the 1-D case. Consequently, δ-operator based implementations of m-D systems cannot be implemented limit cycle free in FXP.

Task T3: 2-D and m-D δ-system models

Discrete-time systems implemented using the δ-operator, as is clear from the discussion above, exhibit superior finite wordlength properties with FLP arithmetic. In the case of FXP arithmetic, they still provide superior coefficient sensitivity. The development of 2-D and m-D models applicable for δ-operator implementations was hence motivated with the

expectation that these properties would still hold true.

The conclusions drawn from the work conducted for task T3 may be summarized as follows: Similar to the 1-D case, under FLP arithmetic, the δ-operator implementation of 2-D and m-D discrete-time systems provides the best choice. Again, this is particularly true in high-order and high-speed applications.

State-space models. In Roesser local s.s. model of q-operator formulated 2-D discrete-time systems takes the form

$$\begin{aligned}
\begin{bmatrix} q_h[\mathbf{x}^h](i,j) \\ q_v[\mathbf{x}^v](i,j) \end{bmatrix} &= \begin{bmatrix} A_q^{(1)} & A_q^{(2)} \\ A_q^{(3)} & A_q^{(4)} \end{bmatrix} \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} + \begin{bmatrix} B_q^{(1)} \\ B_q^{(2)} \end{bmatrix} \mathbf{u}(i,j) \\
&\doteq [A_q] \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} + [B_q]\mathbf{u}(i,j); \\
\mathbf{y}(i,j) &= \begin{bmatrix} C_q^{(1)} & C_q^{(2)} \end{bmatrix} \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} + [D_q]\mathbf{u}(i,j) \\
&\doteq [C_q] \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} + [D_q]\mathbf{u}(i,j),
\end{aligned} \tag{T3.1}$$

where $A_q^{(1)}$ is of size $n_h \times n_h$, $A_q^{(4)}$ is of size $n_v \times n_v$, etc. Also, $q_h[\cdot]$ and $q_v[\cdot]$ denote the horizontal and vertical shift operators, that is,

$$q_h[\mathbf{x}](i,j) = \mathbf{x}(i+1,j) \quad \text{and} \quad q_v[\mathbf{x}](i,j) = \mathbf{x}(i,j+1). \tag{T3.2}$$

To exploit the advantages of δ-operator implementations, analogous to the 1-D case, we define the operators

$$\begin{aligned}
\delta_h[\mathbf{x}](i,j) &= \frac{\mathbf{x}(i+1,j) - \mathbf{x}(i,j)}{\Delta_h} = \frac{q_h[\mathbf{x}](i,j) - \mathbf{x}(i,j)}{\Delta_h}; \\
\delta_v[\mathbf{x}](i,j) &= \frac{\mathbf{x}(i,j+1) - \mathbf{x}(i,j)}{\Delta_v} = \frac{q_v[\mathbf{x}](i,j) - \mathbf{x}(i,j)}{\Delta_v},
\end{aligned} \tag{T3.3}$$

where Δ_h and Δ_v are two positive real constants. The corresponding δ-operator s.s. model may then be obtained as

$$\begin{aligned}
\begin{bmatrix} \delta_h[\mathbf{x}^h](i,j) \\ \delta_v[\mathbf{x}^v](i,j) \end{bmatrix} &= \begin{bmatrix} A^{(1)} & A^{(2)} \\ A^{(3)} & A^{(4)} \end{bmatrix} \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} + \begin{bmatrix} B^{(1)} \\ B^{(2)} \end{bmatrix} \mathbf{u}(i,j) \\
&\doteq [A] \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} + [B]\mathbf{u}(i,j); \\
\mathbf{y}(i,j) &= \begin{bmatrix} C^{(1)} & C^{(2)} \end{bmatrix} \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} + [D]\mathbf{u}(i,j) \\
&\doteq [C] \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} + [D]\mathbf{u}(i,j).
\end{aligned} \tag{T3.4}$$

This is the 2-D version of the intermediate equation mentioned earlier. In addition, as for the 1-D case, we have the following update equations:

$$\begin{aligned} q_h[\mathbf{x}^h](i,j) &= \mathbf{x}^h(i,j) + \Delta_h \cdot \delta_h[\mathbf{x}^h](i,j); \\ q_v[\mathbf{x}^v](i,j) &= \mathbf{x}^v(i,j) + \Delta_v \cdot \delta_v[\mathbf{x}^v](i,j). \end{aligned} \tag{T3.5}$$

Note that,

$$\begin{aligned} A_q = I + \boldsymbol{\Delta} \cdot A_\delta &\iff A_\delta = \boldsymbol{\Delta}^{-1} \cdot (A_q - I_n); \\ B_q = \boldsymbol{\Delta} \cdot B &\iff B_\delta = \boldsymbol{\Delta}^{-1} \cdot B_q; \\ C_q = C_\delta &\iff C_\delta = C_q; \\ D_q = D_\delta &\iff D_\delta = D_q. \end{aligned} \tag{T3.6}$$

Here, $\boldsymbol{\Delta} = [\Delta_h I_{n_h} \oplus \Delta_v I_{n_v}]$ is of size $(n_h + n_v) \times (n_h + n_v)$.

The associated system theoretic notions, such as, transition matrix, transfer function, characteristic equation, etc., have also been introduced. This s.s. model is the basis for designing 2-D filters with superior finite wordlength properties. The design procedures developed are expected to be extremely useful in obtaining high-Q 2-D and m-D digital filters that are suitable for high-speed applications.

Stability. In the 1-D case, it has been shown that, direct techniques with no recourse to transformations (that first converts a given δ-system to its q-system counterpart) can provide numerically more reliable stability checking algorithms. With this in mind, for the 2-D case, a direct stability checking technique applicable to the corresponding δ-system transfer function has been introduced. For this purpose, a recently developed tabular form was extended to the complex coefficient case and the notion of Schur-Cohn minors was introduced to the δ-operator case.

Gramians and balanced realization. The notions of reachability and observability gramians and balanced realization have been introduced for the δ-operator case. In order to do this, first, the relationship between the gramians for the δ- and q-operator cases, as defined in the literature, was established. The reachability and controllability gramians, that is, P and Q, respectively, for 1-D δ-systems were found to satisfy

$$\begin{aligned} P &= \frac{1}{2\pi j} \oint_{\mathcal{T}_\delta} (cI - A_\delta)^{-1} B_\delta B_\delta^* (c^* I - A_\delta^*)^{-1} \frac{dc}{1 + \Delta c}; \\ Q &= \frac{1}{2\pi j} \oint_{\mathcal{T}_\delta} (c^* I - A_\delta^*)^{-1} C_\delta^* C_\delta (cI - A_\delta)^{-1} \frac{dc}{1 + \Delta c}, \end{aligned} \tag{T3.7}$$

where $\mathcal{T}_\delta$ is the stability boundary applicable for δ-systems, that is, $\mathcal{T}_\delta = \{c \in \Im : |c + 1/\Delta| = 1/\Delta\}$. An extension of this is then used to define the 2-D gramians of δ-systems represented in the Roesser model developed above.

For the important class of separable (that is, separable-in-denominator) systems, it is shown that these gramians may be computed through the solution of four Lyapunov equations. These notions and results are useful in many applications, such as, in extracting reduced order models of δ-systems.

Sensitivity. Measures that indicate coefficient sensitivity of the δ-models developed above have been introduced. Unlike what is available in literature, this development is applicable to the MIMO case as well. With these sensitivity measures as a guide, development of minimum sensitivity structures has been carried out. The connection with the corresponding balanced realizations has been pointed out.

Roundoff noise. With the use of a noise model that takes into account the roundoff error propagation in the s.s. model developed above, structures that minimize roundoff noise have been developed.

Publications: Work directly related to grants

[1] K. Premaratne and P.H. Bauer (1994). Limit cycles and asymptotic stability of delta-operator systems in fixed-point arithmetic. *Proceedings 1994 IEEE International Symposium on Circuits and Systems (ISCAS'94),* London, UK, vol. 2, 461-464.

[2] P.H. Bauer and K. Premaratne (1994). Fixed-point implementation of multi-dimensional delta-operator formulated discrete-time systems: Difficulties in convergence. *Proceedings of the 1994 IEEE SOUTHEASTCON,* Miami, FL, 26-29.

[3] K. Premaratne and A.S. Boujarwah (1994). An algorithm for stability determination of two-dimensional delta-operator formulated discrete-time systems. *Multidimensional Systems and Signal Processing,* to appear.

[4] K. Premaratne, J. Suarez, M.M. Ekanayake, and P.H. Bauer (1994). Two-dimensional delta-operator formulated discrete-time systems: State-space realization and its finite wordlength properties. *37th Midwest Symposium on Circuits and Systems,* Lafayette, LA, to be presented; *IEEE Transactions on Signal Processing,* in preparation.

[5] E.C. Kulasekere, K. Premaratne, P.H. Bauer, and L.J. Leclerc (1994). An exhaustive search algorithm for checking limit cycle behavior of digital filters. *IEEE Transactions on Signal Processing,* in preparation.

Note. The contents of [1] and [2] are also being prepared for possible publication in *IEEE Transactions on Signal Processing.*

Publications: Other work where grants are acknowledged

[1] K. Premaratne and E.I. Jury (1994). Discrete-time positive-real lemma revisited: The discrete-time counterpart of the Kalman-Yakubovitch lemma. *IEEE Transactions on Circuits and Systems—I. Fundamental Theory and Applications,* to appear.

[2] M.M. Ekanayake and K. Premaratne (1994). Two-channel IIR QMF filter banks with approximately linear-phase analysis and synthesis filters. *28th Annual Asilomar Conference on Signals, Systems, and Computers,* Pacific Grove, CA, to be presented; *IEEE Transactions on Signal Processing,* in review.

[3] K. Premaratne and M. Mansour (1994). Robust stability of time-variant discrete-time systems with bounded parameter perturbations. *IEEE Transactions on Circuits and Systems—I. Fundamental Theory and Applications,* in review.

[4] S.A. Yost and P.H. Bauer (1994). Robust stability of multi-dimensional difference equations with shift-variant coefficients. *Multidimensional Systems and Signal Processing,* to appear.

Progress Report, September 26, 1994

PROJECT TITLE:

High-speed fixed- and floating-point implementation of delta-operator formulated discrete-time systems

PRINCIPAL INVESTIGATORS:

• **Kamal Premaratne**

University of Miami.

Grant No: N00014-94-1-0454; R&T Project Code: 3148508—01.

• **Peter H. Bauer**

University of Notre Dame.

Grant No: N00014-94-1-0387; R&T Project Code: 3148509—01.

SUMMARY OF PHASE P2 RESULTS

The work described in this report is related to the following

[T2] Task T2: Analysis of nonlinear circuits through δ-operator based schemes.

Problems Posed in Task T2:

Regarding the proposals associated with the above grants, within Task T2, the following questions were raised:

1. With the desirable properties of δ-systems applicable to linear systems in mind, does the same carry over if nonlinear systems are implemented with δ-operator based schemes?

2. In particular, issues concerning coefficient sensitivity and quantization noise is of special importance in such systems.

3. If a δ-operator based scheme offers significant improvements over its q-operator counterpart, the consequences in nonlinear signal processing, nonlinear control, and digital simulation of nonlinear dynamics can be significant.

In fact, the superior finite wordlength performance of the discrete simulation of Chua's Circuit in the grant proposals using the δ-operator, instead of the more conventional q-operator, provided the impetus for the work proposed in Task T2. The work described herein justifies our preliminary optimism and show that this superior performance can be expected with δ-operator based implementations.

This task was proposed to be carried out during Phase P2 with close collaboration between the two PI's. During the whole project duration, both PI's have been in constant contact. In particular, a considerable portion of the work described herein was seen to maturity during a one-week research stay at University of Notre Dame during August 09-16, 1994. During this time, important results that address coefficient sensitivity and quantization error bounds applicable to δ-operator based implementation of nonlinear systems were developed. A description of those Phase P2 results pertaining to coefficient sensitivity follows.

Task T2: Results Pertaining to Coefficient Sensitivity—Summary

Briefly, conclusions drawn from this work may be summarized as follows: We have investigated orbits of linear and nonlinear systems. Several important types of nonlinearities—$\mathcal{C}^1$ nonlinearities, piecewise $\mathcal{C}^1$ nonlinearities, and piecewise linear—were looked into.

- The Fixed-Point Arithmetic (FXP) Case:

With small step size, δ-systems provide superior coefficient sensitivity performance.

- The Floating-Point Arithmetic (FLP) Case:

Conditions under which δ-systems provide superior coefficient sensitivity were derived. Typical digital equivalents of nonlinear systems derived for simulation purposes in fact routinely satisfy these conditions when the step size is small.

Task T2: Results Pertaining to Coefficient Sensitivity—Brief Description

Consider the following q-operator based implementation of a nonlinear system:

$$q[\mathbf{x}](n) = \mathbf{f}_q(\mathbf{x}(n), \mathbf{a}_q), \tag{1}$$

where $q[\mathbf{x}] = \mathbf{x}(n+1)$. Here, $\mathbf{x}(n)$ is the state $\mathbf{x} \in \Re^m$ at time instant n and $\mathbf{a}_q = [a_{q_1}, \ldots, a_{q_M}]^T \in \Re^M$ refers to the system parameters that are *actually stored* within the computer.

The corresponding δ-operator based scheme of the same nonlinear system is of the form

$$\begin{aligned} \delta[\mathbf{x}](n) &= \mathbf{f}_\delta(\mathbf{x}(n), \mathbf{a}_\delta) \quad \text{(Intermediate equation)} \\ q[\mathbf{x}](n) &= \mathbf{x}(n) + \Delta \cdot \delta[\mathbf{x}](n) \quad \text{(Update equation)} \end{aligned} \tag{2}$$

where $\delta[\mathbf{x}](n) = (q[\mathbf{x}](n) - \mathbf{x}(n))/\Delta$ and

$$\mathbf{f}_\delta(\mathbf{x}(n), \mathbf{a}_\delta) = \frac{\mathbf{f}_q(\mathbf{x}(n), \mathbf{a}_q) - \mathbf{x}(n)}{\Delta}. \tag{3}$$

Here, $\Delta \in \Re$ is an arbitrary positive real parameter and $\mathbf{a}_\delta = [a_{\delta_1}, \ldots, a_{\delta_M}]^T \in \Re^M$ again refers to the system parameters that are *actually stored* within the computer.

To see the relationship between $\mathbf{a}_q$ and $\mathbf{a}_\delta$, let the i-th equation in (1) be

$$q[x_i](n) = f_{q_i}(x_1(n), \ldots, x_m(n), a_{q_1}, \ldots, a_{q_M}),\ i = 1, \ldots, m. \tag{4}$$

Then, we may encounter one of two situations:

1. There is a linear term corresponding to x_i, that is, a term of the nature $a_K x_i(n)$, on the RHS of (4). Then, we need to store

$$a_{\delta_i} = \begin{cases} \frac{a_{q_i}}{\Delta}, & \text{for } i = \{1, \ldots, M\} \setminus K; \\ \frac{a_{q_K} - 1}{\Delta}, & \text{for } i = K. \end{cases} \tag{5}$$

2. There is no linear term corresponding to x_i on the RHS of (4). Then, we need to store

$$a_{\delta_i} = \frac{a_{q_i}}{\Delta}, \text{ for } i = \{1, \ldots, M\}, \quad \text{and} \quad \frac{1}{\Delta}. \tag{6}$$

Remark.

1. Of course, in an infinite wordlength implementation, there simply is no difference

between the q- and δ-operator based schemes in (1) and (2). In fact, the latter requires a modest increase in the number of computations. However, what we address is the performance under finite wordlength high-speed conditions.

2. Discretization of a nonlinear system of the form

$$\mathbf{x}^{(1)}(t) = \mathbf{f}(\mathbf{x}(t), \mathbf{a}) \tag{7}$$

can give rise to equations of the type in (1) and (2). Here, $\mathbf{x}^{(i)}$ is the i-th derivative of $\mathbf{x}$.

3. In what follows, $\mathbf{f}(\mathbf{x}, \mathbf{a}) \in \mathcal{C}^1$ denotes a nonlinear function that possesses first partial derivatives.

Now, which of the schemes (1) or (2) yield superior coefficient sensitivity properties of its orbit with respect to perturbations of $\mathbf{a}_q$ or $\mathbf{a}_\delta$, respectively? This consideration is crucial in high-speed applications where a shorter wordlength is the avenue of choice.

In what follows, the following standing assumptions are made:

1. All perturbations are small.

2. Comparison between q- and δ-operator based implementations are done with respect to upper bounds (constructed through appropriate norms) on possible errors due to coefficient sensitivity.

FXP CASE

In the FXP case, a good indication of the coefficient sensitivity of the orbit $\mathbf{x}$ is its first partial derivative with respect to the stored coefficient vector $\mathbf{a}$, that is,

$$\left.\frac{\partial \mathbf{x}}{\partial \mathbf{a}}\right|_n \doteq \frac{\partial}{\partial \mathbf{a}}\mathbf{x}(n) \in \Re^{mM}. \tag{8}$$

I. $\mathcal{C}^1$ nonlinear system

q-operator case

For this case, we can show the following

THEOREM 1. For the q-operator based implementation in (1),

$$\left.\frac{\partial \mathbf{x}}{\partial \mathbf{a}_q}\right|_{n+1} = \sum_{j=0}^{n-1}\left(I_M \otimes \prod_{i=j+1}^{n}\left[\begin{array}{ccc}\frac{\partial \mathbf{f}_q}{\partial x_1} & \cdots & \frac{\partial \mathbf{f}_q}{\partial x_m}\end{array}\right]\Big|_i\right)\cdot \left.\frac{\partial \mathbf{f}_q}{\partial \mathbf{a}_q}\right|_j + \left.\frac{\partial \mathbf{f}_q}{\partial \mathbf{a}_q}\right|_n,$$

where I_M denotes the identity matrix of size $M \times M$ and

$$\left[\begin{array}{ccc}\frac{\partial \mathbf{f}_q}{\partial x_1} & \cdots & \frac{\partial \mathbf{f}_q}{\partial x_m}\end{array}\right]\Big|_i \doteq \left[\begin{array}{ccc}\frac{\partial}{\partial x_1}\mathbf{f}_q(i) & \cdots & \frac{\partial}{\partial x_m}\mathbf{f}_q(i)\end{array}\right] \in \Re^{m\times m};$$
$$\left.\frac{\partial \mathbf{f}_q}{\partial \mathbf{a}_q}\right|_j \doteq \frac{\partial}{\partial \mathbf{a}_q}\mathbf{f}_q(j) \in \Re^{mM}.$$

δ-operator case

For brevity, we only consider the case in (5). Note that,

$$\left.\frac{\partial \mathbf{x}}{\partial \mathbf{a}_\delta}\right|_n = \left.\frac{\partial \mathbf{x}}{\partial \mathbf{a}_q}\right|_n \cdot \Delta. \tag{9}$$

In addition, we need to consider the sensitivity of the orbit with respect to Δ (due to the update equation in (2)). However, if we assume an exact FXP representation for Δ, this term could be ignored.

Using, for instance, a norm to compare the sensitivity measures, we conclude that, the δ-operator based implementation will provide superior coefficient sensitivity performance.

Remark. We obtain similar results when considering the case in (6). Here, one may need to consider sensitivity with respect to $1/\Delta$ as well. Again, we may assume that $1/\Delta$ (and Δ) have exact FXP representations. Even if this is not the case, δ-system is still likely to be superior since the reduction in sensitivity gained through other terms is Δ-fold.

II. Linear system

The superior coefficient sensitivity of the frequency response of δ-operator based systems is thoroughly investigated in Li and Gevers (1990). However, no result exists that address the coefficient sensitivity of the *orbit.*

q-operator case

With the more general result in Theorem 1, we can show the following

THEOREM 2. For the q-operator based implementation $\mathbf{x}(n+1) = A_q\mathbf{x}(n)$, $A_q \in \Re^{m\times m}$,

$$\frac{\partial \mathbf{x}}{\partial A_q}\Big|_{n+1} = \sum_{j=0}^{n-1}(I_m \otimes A_q^{n-j})\overline{U}_{m\times m}(I_m \otimes \mathbf{x})\Big|_j + \overline{U}_{m\times m}(I_m \otimes \mathbf{x})\Big|_n,$$

where $\overline{U}_{q\times p} = \sum_{i=1}^{q}\sum_{j=1}^{p} E_{ij}^{(q\times p)} \otimes E_{ij}^{(q\times p)} \in \Re^{q^2\times p^2}$, $E_{ij}^{(q\times p)} = \mathbf{e}_i^{(q)}\mathbf{e}_j^{(p)^T} \in \Re^{q\times p}$. Here, $\mathbf{e}_i^{(n)} \in \Re^n$ is the unit vector with 1 on its i-th row (Brewer 1978).

δ-operator case

The corresponding δ-system's intermediate equation is $\delta[\mathbf{x}](n) = A_\delta\mathbf{x}(n)$ where $A_\delta = (A_q - I)/\Delta$. The update equation of course is as in (2).

Again, as in Section I, we can show that, the δ-operator based implementation will provide superior coefficient sensitivity performance.

III. Piecewise $\mathcal{C}^1$ nonlinear system

Consider a nonlinearity that is piecewise and possesses first partial derivatives within each 'piece'. To address its coefficient sensitivity, we model the dynamics of such a system as follows:

1. Within each 'piece', the system dynamics is a $\mathcal{C}^1$ nonlinearity.

2. Each instant of the orbit's 'entry' into another 'piece' is modeled as a perturbation in the initial conditions.

Regarding item 1, as previous results indicate, the δ-operator based implementation will be superior within each 'piece'. Regarding item 2, we need to investigate the orbit's coefficient sensitivity with respect to initial conditions. This is addressed now.

q-operator case

A reasonable sensitivity measure is

$$\frac{\partial \mathbf{x}}{\partial \mathbf{x}(0)}\Big|_n \doteq \frac{\partial}{\partial \mathbf{x}(0)}\mathbf{x}(n) \in \Re^{m^2}. \tag{10}$$

Then, we can show the following

THEOREM 3. For the q-operator based implementation in (1),

$$\frac{\partial \mathbf{x}}{\partial \mathbf{x}(0)}\Big|_{n+1} = I_m \otimes \prod_{i=0}^{n} \left[\frac{\partial \mathbf{f}_q}{\partial x_1} \quad \cdots \quad \frac{\partial \mathbf{f}_q}{\partial x_m} \right] \Big|_i$$

δ-operator case

One may show that, Theorem 3 is equally applicable for the δ-operator case as well.

Hence, regarding sensitivity due to initial conditions, both q- and δ-operator based implementations are expected to be provide comparable results.

This implies that, in totality, δ-operator based implementations will provide superior results.

IV. Piecewise linear system

Again, we address the coefficient sensitivity of the orbit with respect to the initial conditions.

q-operator case

As in Section II, with the more general result in Theorem 3, we can show the following

THEOREM 4. For the q-operator based implementation of $\mathbf{x}(n+1) = A_q \mathbf{x}(n)$,

$$\frac{\partial \mathbf{x}}{\partial \mathbf{x}(0)}\Big|_{n+1} = A_q^{n+1}.$$

δ-operator case

Again, one may show that, Theorem 4 is equally applicable for the δ-operator case as well.

Hence, as in Section III, δ-operator based implementations will provide superior results.

FLP CASE

In the FLP case, representable values are spaced farther apart at higher values of the parameter. Hence, instead of that used for the FXP case (see (8)), a more realistic sensitivity measure is (see Li and Gevers (1990))

$$\left.\frac{\partial \mathbf{x}}{\partial \mathbf{a}/\mathbf{a}}\right|_n \doteq \begin{bmatrix} \frac{\partial}{\partial a_1/a_1}\mathbf{x}(n) \\ \vdots \\ \frac{\partial}{\partial a_M/a_M}\mathbf{x}(n) \end{bmatrix} \in \Re^{mM}. \tag{11}$$

I. $\mathcal{C}^1$ nonlinear system

q-operator case

For this case, we can show the following

THEOREM 5. For the q-operator based implementation in (1),

$$\left.\frac{\partial \mathbf{x}}{\partial \mathbf{a}_q/\mathbf{a}_q}\right|_{n+1} = \sum_{j=0}^{n-1} \left(I_M \otimes \prod_{i=j+1}^{n} \left[\frac{\partial \mathbf{f}_q}{\partial x_1} \quad \cdots \quad \frac{\partial \mathbf{f}_q}{\partial x_m} \right] \Big|_i \right) \cdot \left.\frac{\partial \mathbf{f}_q}{\partial \mathbf{a}_q/\mathbf{a}_q}\right|_j + \left.\frac{\partial \mathbf{f}_q}{\partial \mathbf{a}_q/\mathbf{a}_q}\right|_n.$$

δ-operator case

Again, we only consider the case in (5). Also, let us assume that, the elements in $\mathbf{a}_q$ are enumerated such that, for each $i = 1, \ldots, m$, a_i is the 'linear' element of the i-th equation. Then, note that,

$$\begin{aligned} \left.\frac{\partial \mathbf{x}}{\partial \mathbf{a}_\delta/\mathbf{a}_\delta}\right|_n &= \left[\begin{array}{c} a_{\delta_1}\frac{\partial \mathbf{x}}{\partial a_{\delta_1}} \\ \vdots \\ a_{\delta_M}\frac{\partial \mathbf{x}}{\partial a_{\delta_M}} \end{array} \right] \Big|_n \\ &= \left[\begin{array}{c} (a_{q_1}-1)\frac{\partial \mathbf{x}}{\partial a_{q_1}} \\ \vdots \\ (a_{q_m}-1)\frac{\partial \mathbf{x}}{\partial a_{q_m}} \\ a_{q_{m+1}}\frac{\partial \mathbf{x}}{\partial a_{q_{m+1}}} \\ \vdots \\ a_{q_M}\frac{\partial \mathbf{x}}{\partial a_{q_M}} \end{array} \right] \Big|_n, \end{aligned} \tag{12}$$

where we have used (5) and (9). As before, we may ignore the effect of Δ.

Again, we use a norm to compare the sensitivity measures. For instance, using the 1- or ∞-norm, we conclude that, the δ-operator based implementation will provide superior

coefficient sensitivity performance if

$$|a_{q_i} - 1| \leq |a_{q_1}|, \ \forall i = 1, \ldots, m. \tag{13}$$

But, how practical is this restriction? In other words, how often, if at all, is it satisfied in practice? To address this, consider the following

Example. Lorenz equation. Consider the state-space description of the Lorenz equation:

$$\begin{aligned} x_1^{(1)}(t) &= -\sigma\,(x_1(t) - x_2(t)); \\ x_2^{(1)}(t) &= \rho\, x_1(t) - x_2(t) - x_1(t)x_3(t); \\ x_3^{(1)}(t) &= x_1(t)x_2(t) - \beta\, x_3(t). \end{aligned}$$

For digital simulation of the corresponding orbit, we use the forward Euler scheme with step size Δ. This yields

$$\begin{aligned} x_1(n+1) &= (1 - \Delta\sigma)\, x_1(n) + \Delta\sigma\, x_2(n); \\ x_2(n+1) &= \Delta\rho\, x_1(n) + (1 - \Delta)\, x_2(n) - \Delta\, x_1(n)x_3(n); \\ x_3(n+1) &= (1 - \Delta\beta)\, x_3(n) + \Delta\, x_1(n)x_2(n). \end{aligned}$$

We at once observe the following: For a small step size Δ,

1. Linear terms are close to 1.

2. Other terms are very small.

Hence, the condition in (13) is in fact satisfied!

In fact, when digital simulation of nonlinear systems are carried out, (13) is often satisfied for a small step size (which denotes fast sampling). Hence, we conclude that, a δ-operator based implementation of such a simulation will provide superior coefficient sensitivity performance!!

II. Linear system

Again, no result that addresses coefficient sensitivity of the *orbit* of linear systems implemented using FLP arithmetic is available.

Without delving into much detail, we simply state the relevant result: Consider the q-operator based implementation $\mathbf{x}(n+1) = A_q\mathbf{x}(n)$ and its corresponding δ-operator based implementation. With respect to the FLP coefficient sensitivity measure introduced

above, the coefficient sensitivity of the δ-system is superior (in terms of the norm being used) than that for the corresponding q-system if

$$\|A_q - I\| \leq \|A_q\|. \tag{14}$$

It is not hard to show the following:

$$|\lambda_i[A_q] - 1| \leq |\lambda_i[A_q]|, \ \forall i = 1, \ldots, m \iff \|A_q - I\|_F \leq \|A_q\|_F; \tag{15a}$$

$$|\lambda_i[A_q] - 1| \leq |\lambda_j[A_q]|, \ \forall i, j = 1, \ldots, m \iff \|A_q - I\|_2 \leq \|A_q\|_2. \tag{15b}$$

$$|\mathrm{diag}_i[A_q] - 1| \leq |\mathrm{diag}_i[A_q]|, \ \forall i = 1, \ldots, m \iff \|A_q - I\|_{1,\infty} \leq \|A_q\|_{1,\infty}. \tag{15c}$$

Here, $\lambda_i[A_q]$ denotes the i-th eigenvalue of A_q and $\mathrm{diag}_i[A_q]$ denotes the i-th diagonal element of A_q.

Hence, if any one of the above conditions are satisfied, the δ-operator based implementation will provide superior coefficient sensitivity performance.

Remark. Li and Gevers (1993) refers to the region in condition (15a) as the *Middleton-Goodwin (MG) Region.* They have shown that, if (15a) is satisfied, the δ-operator based implementation will provide superior coefficient sensitivity of its *frequency response.*

Regarding systems corresponding to those in Sections III and IV of the FXP case, δ-systems offer similar advantages.

Example, continued. Lorenz equation. To justify and validate the results above, a digital simulation of the Lorenz equation was carried out using both q- and δ-operator schemes with FLP. The results are summarized in the series of graphs.

1. Nominal coefficient values: $\sigma = 10$; $\rho = 28$; $\beta = 8/3$.
2. Initial conditions: $[x_1(0), x_2(0), x_3(0)]^T = [0, 5, 75]^T$.
3. Coefficient representation: FLP arithmetic with the number of bits used for the mantissa indicated on each graph.
4. Integration scheme: Forward Euler with step size $\Delta = 1e - 03$.
5. Number of time steps is 25,000.
6. Only projection onto (x_1, x_2)-plane is shown.

It is important to note that, when only 4 bits are allowed on the mantissa, the qualitative behavior of the q-system is completely different than what is expected. Hoever, the δ-system still provides satisfactory results. Hence, one may use a shorter wordlength for coefficient representation with the latter without affecting performance. The implications on speed, number of components, cost, reliability, etc., are obvious.

References

J.W. Brewer (1978). Kronecker products and matrix calculus in system theory. *IEEE Trans. Circ. Syst.,* **CAS-25**, 772-781.

G. Li and M. Gevers (1990). Comparative study of finite wordlength effects in shift and delta operator paramaterization. *Proc. IEEE CDC'90,* Honolulu, HI, **2**, 954-959.

Progress Report, September 26, 1994

HIGH SPEED FIXED- AND FLOATING-POINT IMPLEMENTATION OF DELTA-OPERATOR FORMULATED DISCRETE-TIME SYSTEMS

Peter H. Bauer and Kamal Premaratne

Focus: Effects of Quantization Errors in Nonlinear Q- and Δ-Operator Systems

Abstract

Absolute quantization error bounds are constructed for q- and δ-operator implementations of the nonlinear system $\underline{x}_{n+1} = f(\underline{x}_n)$. Various assumptions on the type of the nonlinearity $f(\cdot)$ are made and both fixed and floating point formats are investigated. A comparison between the advantages and disadvantages of the two implementation schemes is introduced. Finally, an outlook concerning future work is given.

I. Absolute Bounds on Quantization Errors

I.1. Nonlinearities of the Polynomial Type

I.1.1. Fixed Point Case

Q-operator case:

- System description:

$$\underline{x}_{n+1} = f(\underline{x}_n), \quad f(\cdot) : \Re^M \rightarrow \Re^M$$

where

$$f(\underline{x}_n) = \begin{pmatrix} f_1(x_1, \cdots, x_M) \\ \vdots \\ f_M(x_1, \cdots, x_M) \end{pmatrix},$$

$$f_j(x_1, \cdots, x_M) = \sum_{i_1=0}^{N_j} \cdots \sum_{i_M=0}^{N_j} a^{(j)}_{i_1,\cdots,i_M} \, x_1^{i_1} \cdot \cdots \cdot x_M^{i_M}$$

$$j = 1, \cdots, M.$$

- Assumptions:
  - single precision (i.e., single length accumulators)
  - quantization step: q

- Computed Orbit:
  $\underline{\hat{x}}(n)$

- **Error model for the computed orbit:**

$$f_j(\hat{x}_1, \cdots, \hat{x}_M) = \sum_{i_1=0}^{N_j} \cdots \sum_{i_M=0}^{N_j} a^{(j)}_{i_1 \cdots i_M} \hat{x}_1^{i} \cdots \hat{x}_M^{i_M} + \mu_j$$

where

$$\mu_j = \sum_i \epsilon^{(1)}_{ji} + \sum_i \epsilon^{(2)}_{ji} + \cdots + \sum_i \epsilon^{(M \cdot N_j)}_{ji}$$

and

$$|\epsilon^{(k)}_{ji}| < kq \text{ (truncation)}$$
$$|\epsilon^{(k)}_{ji}| \leq \frac{k}{2} q \text{ (rounding)}$$

- **If the nonlinearity $f_j(\cdot)$ is known, the number of nonzero terms in the polynomial is known and therefore the number of terms in the summations of ϵ-terms. Hence an absolute bound on μ_j can be constructed:**

$$|\mu_j| \leq C_j q \text{ (truncation)}$$

where $C_j = l_1 + 2l_2 + \cdots + M \cdot N_j l_{M \cdot N_j}$

$l_\nu, \nu = 1, \cdots, MN_j$ **being the number of terms present in the summation** $\Sigma_i \ \epsilon^{(\nu)}_{ji}$.

δ-operator case:

- **system description:**

$$\delta\underline{x}(n) = \frac{f(\underline{x}_n) - \underline{x}_n}{\Delta}, \quad x(n+1) = x(n) + \Delta\delta[x(n)].$$

$$\delta x_j(n) = \sum_{i_1=0}^{N_j} \cdots \sum_{i_M=0}^{N_j} \frac{a^{(j)}_{i_1 \cdots i_M}}{\Delta} x_1^{i_1} \cdot \cdots \cdot x_M^{i_M} - \frac{x_j}{\Delta} =$$

$$= \sum_{i_1=0}^{N_j} \cdots \sum_{i_M=0}^{N_j} b^{(j)}_{i_1 \cdots i_M} x_1^{i_1} \cdot \cdots \cdot x_M^{i_M}$$

where

$$b^{(j)}_{i_1,\cdots,i_M} = \begin{cases} \frac{a^{(j)}_{i_1 \cdots i_M} - 1}{\Delta} & \text{for } (i_1, \cdots, i_M) \ \ s.th. \\ & i_j = 1, \ and \ i_\nu = 0 \text{ for} \\ & \nu = 1, \cdots, M, \ \ \nu \neq j \\ \frac{a^{(j)}_{i_1 \cdots i_M}}{\Delta} & \text{otherwise .} \end{cases}$$

- **Assumptions (same as in q-operator case):**
  - **single precision**
  - **quantization step: q**

- **Computed Orbit:**

$$\hat{\underline{x}}(n)$$

- **Error model for the computed orbit:**

$$\delta[\hat{x}_j] = \sum_{i_1=0}^{N_j} \cdots \sum_{i_M=0}^{N_j} b^{(j)}_{i_1 \cdots i_M} \hat{x}_1^{i_1} \cdot \cdots \cdot \hat{x}_M^{i_M} + \mu_j^{(\delta)}$$

where

$$\mu_j^{(\delta)} = \sum_i \mu_{ji}^{(1)} + \sum_i \mu_{ji}^{(2)} + \cdots + \sum_i \mu_{ji}^{(MN_j)}$$

and

$$|\mu_{ji}^{(k)}| < k \cdot q \text{ (truncation)}$$
$$|\mu_{ji}^{(k)}| \leq \frac{k}{2} \cdot q \text{ (rounding)}$$

- **Upper bound on $\mu_j^{(\delta)}$:**

$$|\mu_j^{(\delta)}| \leq (C_j + 1) \cdot q \quad \text{worst case}$$

where C_j is defined as in the q-operator case.

- **Error in the computation of the next state:**

$$\hat{x}_j(n+1) = \hat{x}_j(n) + \Delta\delta[\hat{x}_j(n)] = \sum_{i_1=0}^{N_j} \cdots \sum_{i_M=0}^{N_j} a^{(j)}_{i_1 \cdots i_M} \hat{x}_1^{i_1} \cdots \hat{x}_1^{i_M}$$
$$+ \Delta \cdot \mu_j^{(\delta)}(n) + \mu_{\Delta j}^{(\delta)}(n)$$

where

$$|\mu_{\Delta j}^{(\delta)}(n)| < q \text{ for truncation}$$
$$|\mu_{\Delta j}^{(\delta)}(n)| < \frac{q}{2} \text{ for rounding}$$

Comparison: δ- vs. q-operator:

Error term bound for the q-operator:

$$|\mu_j| \leq C_j \cdot q \text{ (truncation)}$$

Error term bound for the δ-operator:

$$|\Delta \cdot \mu_j^{(\delta)} + \mu_{\Delta j}^{(\delta)}| \leq (C_j+1)q\Delta + q = q([C_j+1]\Delta+1) \text{ (truncation)}$$

- δ-operator formulation has a smaller absolute error bound for:

$$\frac{C_j - 1}{C_j + 1} > \Delta, \text{ for } j = 1, \cdots, M$$

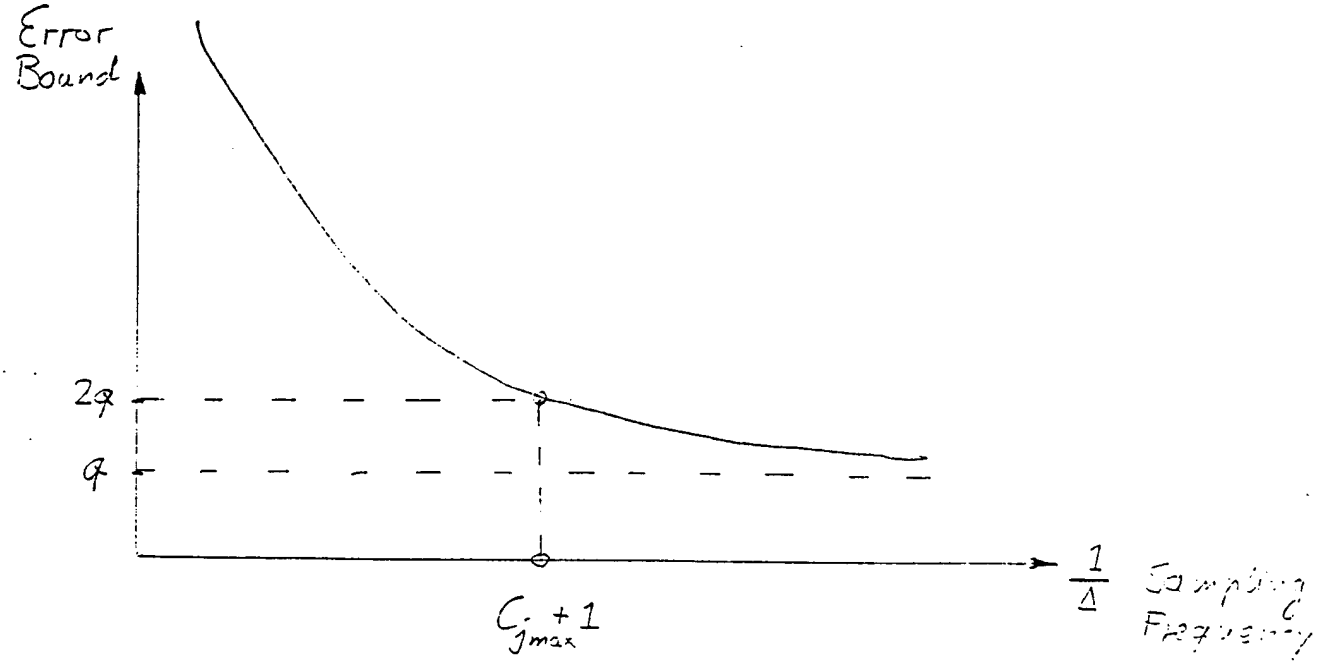
Usually $C_j >> 1$ and hence for

$$1 - \epsilon > \Delta,$$

the δ-operator system is preferable.
(For high speed systems we typically have $\Delta << 1$.)

Reasonable choice of Δ (from an error bound perspective):

$$\Delta \leq \frac{1}{C_{j\max} + 1}$$



Remarks:

- δ-operator implementations in FXP format seem to produce a significantly smaller bound than q-operator implementations, if $\Delta << 1$.
- A δ-operator implementation requires a larger dynamic range than the q-operator implementation, if $\Delta << 1$. $\Rightarrow$ the chance of overflow increases.
- To avoid overflow problems, the δ-operator system needs to be implemented with a larger wordlength.

Forced Response Case:

System description for the q-operator:

$$\begin{aligned} \underline{x}_{n+1} &= f(\underline{x}_n, \underline{u}_n) \\ f(\underline{x}_n, \underline{u}_n) &= \begin{pmatrix} f_1(x_1, \cdots, x_M, u_1, \cdots, u_K) \\ \vdots \\ f_M(x_1, \cdots, x_M, u_1, \cdots, u_K) \end{pmatrix} \end{aligned}$$

where the f_ν's are again multivariate polynomials in up to $M + K$ variables.

System description for the δ-operator:

$$\begin{aligned} \delta[\underline{x}_n] &= \frac{f(\underline{x}_n) - \underline{x}_n}{\Delta} \\ \underline{x}_{n+1} &= \underline{x}_n + \Delta\delta[\underline{x}_n] \end{aligned}$$

- Using a similar error model as in the zero-input case, the computation of $\underline{x}_{n+1}$ in the q-operator case and the computation of $\delta\underline{x}_n$ in the δ-operator case produce bounds of similar magnitude.
- If $\Delta << 1$, the δ-operator system again has an advantage over the q-operator system, since the errors of the first equation are much larger than the ones produced in the update equation.

I.1.2. The Floating Point Case

Q-Operator model – ideal case:

$$x_j(n+1) = \sum_{i_1} \cdots \sum_{i_M} a^{(j)}_{i_1 \cdots i_M} \, x_1^{i_1} \cdots x_M^{i_M} \tag{1}$$

δ-operator model – ideal case:

$$\delta x_j(n) = \sum_{i_1} \cdots \sum_{i_M} \frac{a^{(j)}_{i_1 \cdots i_M}}{\Delta} x_1^{i_1} \cdots x_M^{i_M} - \frac{x_j}{\Delta} \tag{2a}$$

$$x_j(n+1) = x_j(n) + \Delta \delta[x_j(n)] \tag{2b}$$

Model for floating point errors due to multiplication and addition:

$$\begin{aligned} x \odot y &= xy(1+\epsilon) \\ x \oplus y &= x(1+\epsilon) + y(1+\epsilon_2) \end{aligned}$$

Consider two cases:

(a) $a_{i_1,\cdots,i_M} \simeq 1$ with $i_\nu = 0$ for $\nu = 1, \cdots, M, \nu \neq j$
and $i_j = 1$.
$|a_{i_1,\cdots,i_M}| << 1$ for all other combinations of $(i_1, \cdots, i_M)$,
$\underline{x}(n) \epsilon [-1, +1]^M$

(b) condition (a) is not satisfied.

Case (a):

δ-operator bounds on quantization error are much smaller than q-operator bounds.

Qualitative explanation:

For case (a), all partial sums and products in the computation of $\Delta \cdot \delta[x_j(n)]$ are much smaller than $x_j(n)$. Therefore the errors in the computation of $\Delta\delta[x_j(n)]$ are smaller compared to the final addition error in (2b). Therefore the δ-operator model implicitely performs operand sorting, which is known to reduce quantization errors in floating point arithmetic.

Case (b):

δ-operator error bounds are slightly larger than q-operator bounds.

Other classes of nonlinear systems exist which also perform better using a δ-operator formulation. One such class is the weakly nonlinear functions satisfying:

$$\begin{aligned} f_j(x_1, \cdots, x_M) &= x_j + \epsilon_j(x_1, \cdots x_M) \\ &\quad \text{with } |\epsilon_j(x_1, \cdots, x_M)| << |x_j| \\ &\quad j = 1, \cdots, M. \end{aligned}$$

(If $f_j(x_1, \cdots, x_M)$ is of polynomial type, the system has to operate on a hypercuboid or another finite subspace of $\Re^M$ since polynomials cannot be weakly nonlinear in the above sense for all $x_i \epsilon \Re$.)

Note:

If equations (1) or (2) arise from quantizing a continuous time system with a very short sampling time, then the condition

$$|x_j(n)| >> |\Delta\delta x_j(n)|$$

can be satisfied giving the δ-operator formulation an advantage over the q-operator.

I.1.3. A Generalized Delta-Operator Model for Linear and Nonlinear Systems

Linear Case:

Assume the system is given by:

$$\underline{x}(n+1) = A_q\underline{x}(n) + B_q\underline{u}(n)$$

Consider the modified Δ-operator form:

$$\begin{aligned}\delta[x(n)] &= \frac{A_q - A_0}{\Delta}\underline{x}(n) + \frac{B_q - B_0}{\Delta}\underline{u}(n)\\ x(n+1) &= A_0\underline{x}(n) + B_0\underline{u}(n) + \Delta\delta[x(n)]\end{aligned}$$

with A_0 and B_0 being integer matrices closest to A_q and B_q respectively.

Advantages

- **The dynamic range of the A^δ and B^δ matrices becomes smaller and the chance of overflow is reduced.**
- **The delta operator realization has the same improved sensitivity as in the regular delta-operator case.**
- **In floating point arithmetic, the condition $A_q \simeq I$ is not necessary for improved error behavior of the delta-operator system.**

Nonlinear Case:

A similar argument as in the linear case can be made for weakly nonlinear systems of the form:

$$\underline{x}(n+1) = A_q \underline{x}(n) + B_q \underline{u}(n) + \underline{\epsilon}(\underline{x}(n), \underline{u}(n))$$

where

$$\| \underline{\epsilon}(\underline{x}(n), \underline{u}(n)) \| << \| \underline{x}(n) \|$$

and $$\| \underline{\epsilon}(\underline{x}(n), \underline{u}(n)) \| << \| \underline{u}(n) \|$$

I.2. Nonlinearities of Piecewise Linear Form

I.2.1. The Fixed Point Case

Although a piecewise linear continuous scalar function f: $\Re \rightarrow \Re$ can be represented as

$$f(x) = \sum_i (|x - \mu_i| a_i) + b,$$

a computationally more efficient realization is:

$$f(x) = c_i x + d_i \text{ for } \underline{x}_i \leq x \leq \bar{x}_i \tag{1}$$

Therefore, the resulting system $\underline{x}_{n+1} = f(\underline{x}_n)$ can be written in form of several linear state space equations with a driving term, and the driving terms being known *a priori*:

δ – operator:

$$\delta[\underline{x}_n] = A_i^\delta \underline{x}_n + \underline{u}_i^\delta$$
$$\underline{x}_{n+1} = \underline{x}_n + \Delta\delta[\underline{x}_n],\ i = 1, \cdots, K$$

q – operator:

$$\underline{x}_{n+1} = A_i^q \underline{x}_n + \underline{u}_i^q,\ i = 1, \cdots, K$$

Conclusion:

- For single precision (quantization after products) the absolute error bounds for the δ-operator realization are smaller than for the q-operator realization.
- For double precision (quantization after summation) the absolute error bounds for the δ-operator realization are approximately the same as for the q-operator.

I.2.2. The Floating Point Case

Due to (1), the system can be modeled as a time-variant linear system with a known, piecewise constant input. Therefore the same conclusions apply as in the linear t.i.v. case with regard to absolute error bounds:

- Generally, absolute error bounds of the δ- and q-operator system are of similar size.

- If the resulting A-matrices of the piecewise linear system are all 'close' to the identity matrix I, then the δ-operator system will perform superior to the q-operator. (see comments in I.1.2.). This requires that the driving terms are also small relative to the states.

I.3. Sector Bounded Nonlinear Functions

I.3.1. The Fixed Point Case

System description:

$$\begin{aligned} x_i(n+1) &= \mathcal{F}_{i1}[a_{i1}x_1(n)] + \cdots + \mathcal{F}_{im}[a_{im}x_m(n)] \\ i &= 1, \cdots, m. \end{aligned}$$

Sector conditions on $\mathcal{F}[\ ]$:

$$\mathcal{F}_{ij}(x) = k_{ij}x, \ k_{ij} \in [\underline{k}_{ij}, \bar{k}_{ij}].$$

If $\epsilon_{ij}(n)$ is the error affiliated with the computation of $\mathcal{F}_{ij}[\cdot]$, the response of the q and δ-operator system can be absolutely bounded by the following majorant system:

q-operator:

$$x_i^+(n+1) = \sum_{j=1}^{m} m_{ij}^+ x_j^+(n) + \sum_{j=1}^{m} \epsilon_{ij}^+(n), \ \ i = 1, \cdots, m$$

where

$$\begin{aligned} m_{ij}^+ &= \max\{|\underline{k}_{ij}a_{ij}|, |\bar{k}_{ij}a_{ij}|\} \\ \epsilon_{ij}^+(n) &= |\epsilon_{ij}(n)| \end{aligned}$$

δ-operator

$$x_i^+(n+1) = \sum_{j=1}^{m} m_{ij}^+ x_j^+(n) + \Delta\left(\epsilon_\Delta^+(n) + \sum_{j=1}^{m} \epsilon_{ij}^+(n)\right) + \epsilon_{up}^+(n),$$
$$i = 1, \cdots, m$$

where

$\epsilon_{up}^+(n) = |\epsilon_{up}(n)|$, $\epsilon_{up}(n)$: **error occurring in update equation**
$\epsilon_\Delta^+(n) = |\epsilon_\Delta(n)|$, $\epsilon_\Delta(n)$: **error due to division by Δ.**

Comparison:

The bound for the δ-operator implementation is lower if

$$\max\left(\Delta(\epsilon_\Delta^+(n) + \sum_{j=1}^{m} \epsilon_{ij}^+(n)) + \epsilon_{up}^+(n)\right) < \max\left(\sum_{j=1}^{m} \epsilon_{ij}^+(n)\right)$$

- **Since the bound for $|\epsilon_{ij}^+(n)|$ is typically much larger than for $|\epsilon_\Delta^+(n)|$ or $|\epsilon_{up}^+(n)|$, it is obvious that for $\Delta << 1$, the above condition is always satisfied.**
- **A similar result holds if the nonlinearities $\mathcal{F}$ enter the system in a different form, i.e., if they have arguments which consist of partial sums.**
- **A similar comparison arises for other fixed point quantization formats.**

I.3.2. The Floating Point Case

Generally, the δ-operator implementation is not superior to the q-operator implementation if one compares absolute error bounds. However, as stated before (I.1.2, I.2.2), if the condition

$$|x_i(n)| >> |\Delta\delta(x_i(n))|$$

holds true for all states $(i = 1, \cdots, m)$, then the δ-operator implementation has a significantly smaller error bound.

A class of systems satisfying the above condition is given by:

$$\begin{aligned} x_i(n+1) &= \mathcal{F}_{i1}[a_{i1}x_1(n)] + \cdots + \mathcal{F}_{im}[a_{im}x_m(n)] \\ & \quad i = 1, \cdots, m \end{aligned}$$

where

$$\begin{aligned} \mathcal{F}_{ij}(x) &= k_{ij}x, \quad k_{ij} \in [\underline{\epsilon}_{ij}, \overline{\epsilon}_{ij}] \text{ for } i \neq j, \\ & \qquad k_{ii} \in [1 - \epsilon_{ii}, 1 + \epsilon_{ii}] \text{ otherwise,} \\ & \quad |\epsilon_{ij}| << 1 \text{ for } i, j = 1, \cdots, m. \end{aligned}$$

Again, such a system could arise from a continuous time system with a high sampling rate.

II. Comparison of Implementations

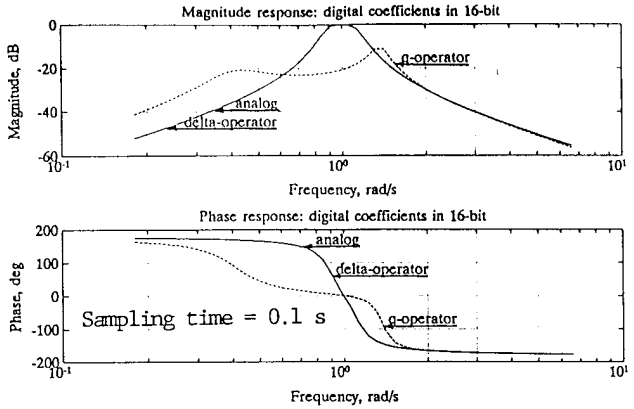
	FXP-case	FLP-case	BFLP-case*
general system: q-error bounds	δ-operator is mostly superior	δ and q-operator are comparable	similar to FXP case?
q-error bounds for a short sampling time in the discretization process	δ-operator is mostly superior	δ operator is superior	similar to FXP case?
limit cycles (linear case only)	δ-operator produces incorrect equilibria	limit cycles in underflow for both q and δ-operator	similar to FLP case
hardware requirements for small Δ	δ-operator requires longer registers than q-operators	independent of Δ	
overflow effects	δ-operator is more likely to cause overflow	in both operators unlikely	similar to FLP case
general sensitivity	δ operator superior	δ-operator better than or equal to q-operator	similar to FXP case?
sensitivity for a short sampling time in discretization process	δ-operator superior	δ operator superior	similar to FXP case?

* has not been analyzed in detail yet, expected results.

HIGH-SPEED FIXED- AND FLOATING-POINT IMPLEMENTATION OF DELTA-OPERATOR FORMULATED DISCRETE-TIME SYSTEMS

PETER H. BAUER, University of Notre Dame (Grant No: N00014-94-1-0387)
KAMAL PREMARATNE, University of Miami (Grant No: N00014-94-1-0454)

FIGURE



FREQUENCY RESPONSE OF AN ANALOG PROTOTYPE AND ITS DIGITAL EQUIVALENTS. DIGITAL COEFFICIENTS ARE STORED WITH 16 BITS IN MANTISSA. (Note. Analog and delta-system plots overlap)

APPROACH

*Limit cycle behavior (deadband bounds).
*Coefficient sensitivity (differential sensitivity measures).
*Quantization error (construction of error envelopes).
*Development of balanced forms for 2-D and *m*-D systems.

The approach is applied to three system types:
[T1] Linear, shift-invariant discrete-time systems.
[T2] Digital simulation of nonlinear systems.
[T3] 2-D and *m*-D discrete-time systems.

OBJECTIVE

Application: High performance, real-time applications involving *fast sampling/short wordlength*.

Conventional shift-operator (*q*-operator) based algorithms are ill-conditioned.

Is the delta-operator (δ-operator) based approach more suitable?

*For which classes of systems does it possess better finite wordlength properties?
*Can it improve reliability of computations in simulating nonlinear and multidimensional systems?

ACCOMPLISHMENTS

*With floating-point, δ-systems offer superior performance (especially, for sampled continuous-time systems) with short step size. All fixed-point δ-systems are plagued by limit cycles.
*δ-operator based digital simulation of nonlinear systems offer superior performance (especially, when using a small discretization step size).
*δ-operator models developed for 2-D and *m*-D filters also possess similar properties.

PRELIMINARIES

OPERATORS

- For $\mathbf{x} \in \Re^m$, $q[\cdot]$ is the operator

$$q[\mathbf{x}](n) = \mathbf{x}(n+1).$$

- For $\mathbf{x} \in \Re^m$, $\delta[\cdot]$ is the operator

$$\delta[\mathbf{x}](n) = \frac{\mathbf{x}(n+1) - \mathbf{x}(n)}{\Delta} = \frac{q[\mathbf{x}](n) - \mathbf{x}(n)}{\Delta}.$$

  Here, Δ is a positive constant (usually the sampling time).
- $q[\cdot]$ and $\delta[\cdot]$ are related by

$$q = 1 + \Delta\delta.$$

q-OPERATOR BASED STATE-SPACE MODEL

q-operator based model $\{A_q, B_q, C_q, D_q\}$ of a linear, shift-invariant, causal, p-input, q-output discrete-time system:

$$\begin{aligned} q[\mathbf{x}](n) &= A_q\mathbf{x}(n) + B_q\mathbf{u}(n); \\ \mathbf{y}(n) &= C_q\mathbf{x}(n) + D_q\mathbf{u}(n). \end{aligned}$$

δ-OPERATOR BASED STATE-SPACE MODEL

Corresponding δ-operator based model $\{A_\delta, B_\delta, C_\delta, D_\delta\}$:

Intermediate equation

$$\begin{aligned} \delta[\mathbf{x}](n) &= A_\delta\mathbf{x}(n) + B_\delta\mathbf{u}(n); \\ \mathbf{y}(n) &= C_\delta\mathbf{x}(n) + D_\delta\mathbf{u}(n). \end{aligned}$$

Update equation

$$q[\mathbf{x}](n) = \mathbf{x}(n) + \Delta \cdot \delta[\mathbf{x}](n).$$

- $\{A_q, B_q, C_q, D_q\}$ and $\{A_\delta, B_\delta, C_\delta, D_\delta\}$ are related by

$$A_q = I + \Delta A_\delta; \quad B_q = \Delta B_\delta; \quad C_q = C_\delta; \quad D_q = D_\delta.$$

[T1] LINEAR, SHIFT-INVARIANT DISCRETE-TIME SYSTEMS

*OBJECTIVE

- How do δ-systems perform under fast sampling/short wordlength conditions? What are their properties regarding limit cycles, quantization errors, coefficient sensitivity, and dynamic range?

*ACCOMPLISHMENTS

Both FXP (fixed-point) and FLP (floating-point) schemes are tackled.

*LIMIT CYCLES

- FXP case: δ-systems (with small Δ) always exhibit limit cycles.
- FLP case: Similar to q-systems, with sufficient mantissa length, limit cycles occur only in underflow.

*QUANTIZATION ERROR PROPAGATION

- FXP case: δ-systems possess smaller bounds for quantization after multiplication. Otherwise, both q- and δ-systems are comparable.
- FLP case: In general, δ-systems are better than or equal to q-systems. If δ-system is the digital equivalent of a continuous-time system with fast sampling, it offers superior performance.

*COEFFICIENT SENSITIVITY

- FXP case: δ-systems are superior with fast sampling.
- FLP case: In general, δ-systems are better than or equal to q-systems. If δ-system is the digital equivalent of a continuous-time system with fast sampling, it offers superior performance.

*DYNAMIC RANGE CONSTRAINTS

- FXP case: If δ-system is the digital equivalent of a continuous-time system, both q- and δ-systems are comparable. If a q-system is simply converted to a δ-system, the latter requires a larger dynamic range.
- FLP case: If δ-system is the digital equivalent of a continuous-time system, it is superior. If a q-system is simply converted to a δ-system, the latter requires a slightly larger dynamic range.

LIMIT CYCLES

The ideal linear system is taken to be asymptotically stable. We consider the zero input case.

FXP Case

A δ-system implementation, under finite wordlength, becomes

$$\delta[\mathbf{x}](n) = Q\{A_\delta \mathbf{x}(n)\};$$
$$q[\mathbf{x}](n) = \mathbf{x}(n) + Q\{\Delta \cdot \delta[\mathbf{x}](n)\}.$$

Here, $Q\{\cdot\}$ is the quantization nonlinearity.

Accomplishments

- δ-systems exhibit DC limit cycles if

$$\Delta \leq 0.5 \quad \text{for rounding;} \quad \Delta < 1.0 \quad \text{for truncation.}$$

  Fundamental reason for these limit cycles is the deadzone of quantizer. This creates deadbands for both $\delta[\mathbf{x}]$ and $\mathbf{x}$.
- In fact, limit cycle free δ-system implementations do not exist!
- A smaller sampling time Δ yields a larger deadband for $\delta[\mathbf{x}]$.
- Construction of this deadband for various arithmetic schemes have been performed.
- Structure of system matrix A_δ has a major effect on geometry of deadband for $\mathbf{x}$.
- Reduction of quantizer deadzone reduces size of deadband, thus reducing DC limit cycle amplitude. But, this increases other (oscillatory) limit cycles.
- Neither the use of unconventional quantization nonlinearities nor scaling techniques overcome this difficulty.

FLP Case

Accomplishments

- If mantissa length is sufficiently large, response will always converge into underflow.
- Hence limit cycles may occur only in underflow. This is usually acceptable if dynamic range of underflow is sufficiently small (that is, smallest representable exponent is sufficiently small).

QUANTIZATION ERROR PROPAGATION

Quantization error propagation is investigated via error envelopes.

FXP Case

Accomplishments

- Error envelopes for δ-systems are lower than for corresponding q-systems if quantization occurs after multiplication. Otherwise, they are comparable.

FLP Case

Accomplishments

- In general, error envelopes δ-systems are better than or equal to q-systems.
- However, when q-system matrix A_q is of the form

$$A_q = I + \{\epsilon_{i,j}\},$$

where the matrix elements $\epsilon_{i,j}$ satisfy

$$| \epsilon_{i,j} | << 1,$$

δ-system provides superior performance. This situation occurs, when a digital equivalent of a continuous-time system is obtained with fast sampling.

- In this situation, δ-operator implementation achieves 'operand sorting' (which is known to tremendously reduce quantization errors in FLP realizations).
- Generalized versions of δ-operator, that can tackle situations where A_q does not satisfy the above condition, have been developed. These provide superior performance than q-systems.

COEFFICIENT SENSITIVITY

Coefficient sensitivity is investigated via differential sensitivity measures. Small perturbations are assumed.

- Frequency response sensitivity have been investigated by others.
- Time response or orbit sensitivity arises as a special case of our work in Task [T2] below.

FXP Case

Accomplishments

- δ-systems offer superior performance, in particular, with fast sampling.

FLP Case

Accomplishments

- In general, δ-systems are better than or equal to the corresponding q-systems.
- Conditions under which δ-systems perform better are derived. In particular, if the δ-system is a digital equivalent of a continuous-time system obtained with fast sampling, it offers superior performance.

DYNAMIC RANGE CONSTRAINTS

FXP Case

Accomplishments

- If the δ-system is obtained by discretization of a continuous-time system, the dynamic range requirements of corresponding q- and δ-systems are comparable.
- If the δ-system is obtained by simply converting a q-system, it typically requires a larger dynamic range, larger coefficient registers, and larger accumulators.

FLP Case

Accomplishments

- Wordlength requirements for q- and δ-systems are comparable.
- If the δ-system is obtained by discretization of a continuous-time system with fast sampling, its zero convergence can be guaranteed with less number of bits.

[T2] DIGITAL SIMULATION OF NONLINEAR SYSTEMS

*OBJECTIVE

- Can one perform *reliable* digital simulations of nonlinear systems using δ-operator based numerical schemes?
- If so, just as for linear systems, would one get superior finite wordlength properties?
- The resulting impact and consequences in high performance computing (for example, in digital simulation of nonlinear systems, signal processing, and control) can be significant.

*ACCOMPLISHMENTS

Several important types of nonlinearities were considered.

*LIMIT CYCLES

This is quite similar to the linear case. See our work in Task [T1].

*QUANTIZATION ERROR PROPAGATION

- FXP case: Due to possibility of incorrect equilibria, FXP implementation is not recommended.
- FLP case: Conditions under which δ-systems are superior are derived.

*COEFFICIENT SENSITIVITY

- FXP case: With small grid size, δ-operator based numerical schemes are superior than the conventional q-operator schemes.
- FLP case: Conditions under which coefficient sensitivity of δ-systems are superior are derived. Typical digital equivalents of nonlinear systems under small grid size routinely satisfy these conditions.

*DYNAMIC RANGE CONSTRAINTS

This is quite similar to the linear case. See our work in Task [T1].

q-OPERATOR BASED NONLINEAR SYSTEM

$$q[\mathbf{x}](n) = \mathbf{f}_q(\mathbf{x}(n), \mathbf{a}_q).$$

- $\mathbf{a}_q = [a_{1_q}, \ldots, a_{M_q}]^T$ are the coefficients that are *actually stored* in computer.

δ-OPERATOR BASED NONLINEAR SYSTEM

We propose the following:

Intermediate equation

$$\delta[\mathbf{x}](n) = \mathbf{f}_\delta(\mathbf{x}(n), \mathbf{a}_\delta).$$

Update equation

$$q[\mathbf{x}](n) = \mathbf{x}(n) + \Delta \cdot \delta[\mathbf{x}](n).$$

- $\delta[\mathbf{x}](n) = (q[\mathbf{x}](n) - \mathbf{x}(n))/\Delta$ and $\mathbf{f}_\delta = (\mathbf{f}_q - \mathbf{x})/\Delta$.
- Δ is an arbitrary positive constant (usually the grid size).
- $\mathbf{a}_\delta = [a_{\delta_1}, \ldots, a_{\delta_M}]^T$ are the coefficients that are *actually stored.*

QUANTIZATION ERROR PROPAGATION

FXP Case

Accomplishments

- δ-systems offer superior performance if quantization is performed after multiplication or if polynomial nonlinearities of higher order are to be implemented.
- However, in FXP, δ-systems may converge to incorrect equilibria (see comments in [T1]). Hence, FXP implementation is not recommended.

FLP Case

Accomplishments

- δ-systems show significantly reduced quantization error bounds if $\delta[\mathbf{x}](n) = \mathbf{f}_\delta(\mathbf{x}(n))$ where $\|\Delta \cdot \mathbf{f}_\delta(\mathbf{x}(n))\| << \|\mathbf{x}(n)\|$.
- Under fast sampling, similar to the linear case, this condition is routinely satsified. Hence, δ-operator based discretization schemes, in FLP, can *drastically reduce* quantizations errors with fast sampling.

COEFFICIENT SENSITIVITY

For this presentation, the nonlinearity is taken to belong to $\mathcal{C}^1$, that is, it possesses first partial derivatives. Small perturbations are assumed.

FXP Case

Coefficient perturbation is approximately independent of its nominal value. Hence, a good sensitivity measure of orbit $\mathbf{x}$ is $\partial\mathbf{x}/\partial\mathbf{a}|_n \doteq \partial\mathbf{x}(n)/\partial\mathbf{a}$.

Accomplishments

- Comparison of q- and δ-systems: $\partial\mathbf{x}/\partial\mathbf{a}_q|_n = \Delta \cdot \partial\mathbf{x}/\partial\mathbf{a}_\delta|_n$. Hence, δ-operator based schemes offer superior coefficient sensitivity when Δ is small.
- Similar comments hold true for linear systems, piecewise $\mathcal{C}^1$ nonlinear systems, and piecewise linear systems.

FLP Case

Coefficient perturbation is approximately proportional to its nominal value. Hence, a good sensitivity measure of orbit $\mathbf{x}$ is

$$\frac{\partial\mathbf{x}}{\partial\mathbf{a}/\mathbf{a}}\Big|_n \doteq \begin{bmatrix} \frac{\partial}{\partial a_1/a_1}\mathbf{x}(n) \\ \vdots \\ \frac{\partial}{\partial a_M/a_M}\mathbf{x}(n) \end{bmatrix}.$$

Accomplishments

- Comparison of q- and δ-systems: We have shown that, δ-operator based schemes offer superior coefficient sensitivity if

$$|a_{i_q} - 1| \leq |a_{i_q}|, \ \forall i = 1, \ldots m.$$

  Here, a_{i_q} indicates the 'linear' term in the i-th equation of $\mathbf{f}_q$.
- Similar comments hold true for linear systems, piecewise $\mathcal{C}^1$ nonlinear systems, and piecewise linear systems.

Example: Lorenz Equation

Consider the digital simulation of Lorenz equation:

$$x_1^{(1)}(t) = a_{11}x_1(t) + a_{12}x_2(t);$$
$$x_2^{(1)}(t) = a_{21}x_1(t) + a_{22}x_2(t) + a_{213}x_1(t)x_3(t);$$
$$x_3^{(1)}(t) = a_{33}x_3(t) + a_{312}x_1(t)x_2(t).$$

Here, $a_{11} = -\sigma$, $a_{12} = \sigma$, $a_{21} = \rho$, $a_{22} = -1$, $a_{213} = -1$, $a_{33} = -\beta$, and $a_{312} = 1$.

q-operator based forward Euler scheme with $\Delta = 1e-04$

$$q[x_{1_q}](n) = a_{11_q}x_{1_q}(n) + a_{12_q}x_{2_q}(n);$$
$$q[x_{2_q}](n) = a_{21_q}x_{1_q}(n) + a_{22_q}x_{2_q}(n) + a_{213_q}x_{1_q}(n)x_{3_q}(n);$$
$$q[x_{3_q}](n) = a_{33_q}x_{3_q}(n) + a_{312_q}x_{1_q}x_{2_q}(n).$$

Here, $a_{11_q} = 1-\Delta\sigma$, $a_{12_q} = \Delta\sigma$, $a_{21_q} = \Delta\rho$, $a_{22_q} = 1-\Delta$, $a_{213_q} = -\Delta$, $a_{33_q} = 1 - \Delta\beta$, and $a_{312_q} = \Delta$.

δ-operator based forward Euler scheme with $\Delta = 1e-04$

$$\delta[x_{1_\delta}](n) = a_{11_\delta}x_{1_\delta}(n) + a_{12_\delta}x_{2_\delta}(n);$$
$$\delta[x_{2_\delta}](n) = a_{21_\delta}x_{1_\delta}(n) + a_{22_\delta}x_{2_\delta}(n) + a_{213_\delta}x_{1_\delta}(n)x_{3_\delta}(n);$$
$$\delta[x_{3_q}](n) = a_{33_\delta}x_{3_\delta}(n) + a_{312_\delta}x_{1_\delta}x_{2_\delta}(n).$$

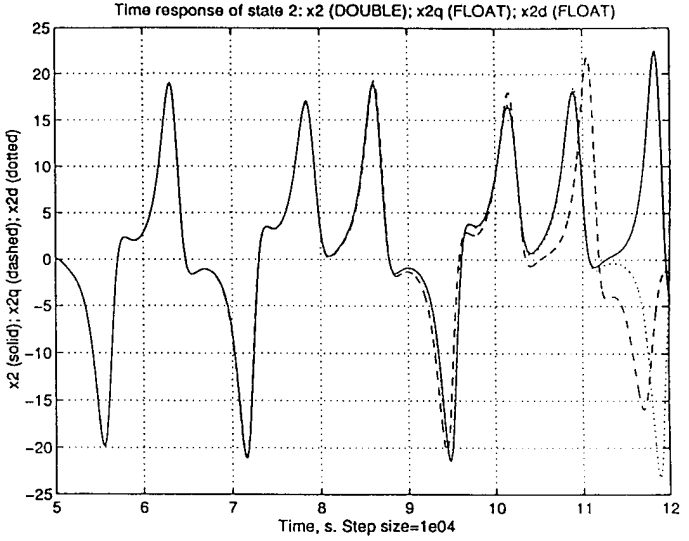
Here, $a_{11_\delta} = a_{11}$, $a_{12_\delta} = a_{12}$, $a_{21_\delta} = a_{21}$, $a_{22_\delta} = a_{22}$, $a_{213_\delta} = a_{213}$, $a_{33_\delta} = a_{33}$, and $a_{312_\delta} = a_{312}$.

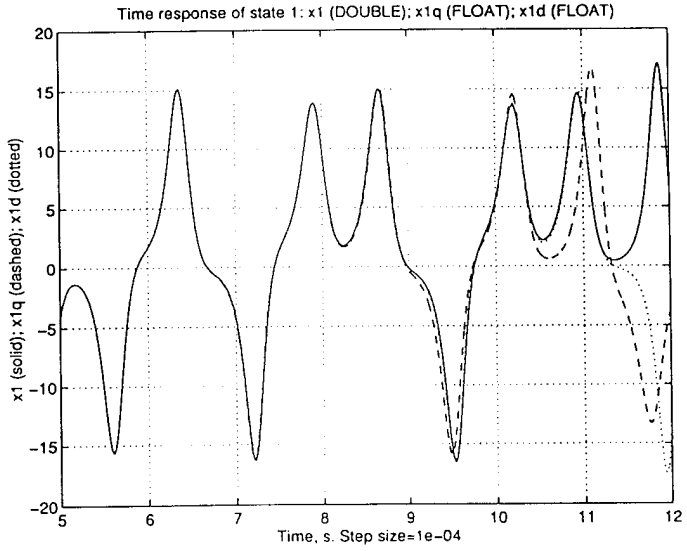
Simulation data

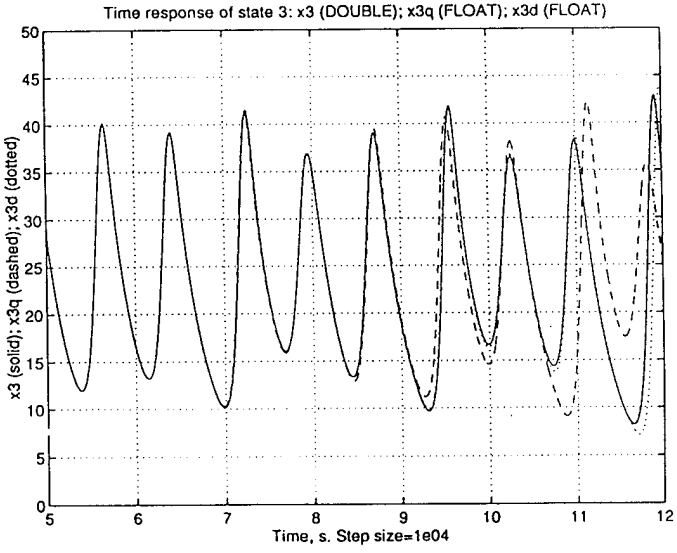
- Nominal coefficient values: $\sigma = 10$; $\rho = 28$; $\beta = 8/3$. This system exhibits chaotic behavior.
- Initial conditions: $\mathbf{x}_q(0) = \mathbf{x}_\delta(0) = [0, 5, 75]^T$.
- Data type: Two simulations were implemented in C using both FLOAT (32-bit FLP) and DOUBLE (64-bit FLP) data types.
- Comparison: DOUBLE simulations until 8 s (where both q- and δ- DOUBLE schemes are identical) were taken as a benchmark for comparison of FLOAT simulations. Clearly, the computed orbit from the δ-scheme is more reliable for a longer duration!

State responses of the Lorenz equation using q- and δ-operator based integration schemes with DOUBLE (64-bit FLP) and FLOAT (32-bit FLP) data types.

1. Nominal coefficient values: $\sigma = 10$; $\rho = 28$; $\beta = 8/3$.
2. Initial conditions: $\mathrm{x}_q(0) = \mathrm{x}_\delta(0) = [0, 5, 75]^T$.
3. Integration scheme: Forward Euler with step size $\Delta = 1e-04$.
4. Coefficients being stored:
   4.1. q-operator based scheme: $\mathrm{x}_q(0)$; a_{11_q}; a_{12_q}; a_{21_q}; a_{22_q}; a_{213_q}; a_{33_q}; a_{312_q}.
   4.2. δ-operator based scheme: $\mathrm{x}_\delta(0)$; a_{11_δ}; a_{12_δ}; a_{21_δ}; a_{22_δ}; a_{213_δ}; a_{33_δ}; a_{312_δ}; Δ.
5. Data type:
   5.1. DOUBLE q- and δ-schemes: Both schemes are identical until approximately 28 s. These are shown as 'solid' lines.
   5.2. FLOAT q-scheme: These are shown as 'dashed' lines.
   5.3. FLOAT δ-scheme: These are shown as 'dotted' lines.







[T3] 2-D AND m-D DISCRETE-TIME SYSTEMS

*OBJECTIVE

- Do the superior finite wordlength properties hold true if 2-D and m-D discrete-time systems are implemented using δ-operator?
- If so, such implementations are useful in high performance, real-time applications that use fast sampling/short wordlength.

*ACCOMPLISHMENTS

*FUNDAMENTAL SYSTEM THEORETIC CONCEPTS

- δ-operator analog of the 2-D Roesser q-model.
- Notions of characteristic equation, transfer function, stability, etc., have been developed.
- Algorithm to check stability, notions of gramians and balanced realizations have been developed.

*COEFFICIENT SENSITIVITY

- FXP case: Balanced realizations possess 'minimum' coefficient sensitivity.
- FLP case: Conditions under which δ-systems perform better are derived. Typically, narrowband high speed digital filters satisfy these requirements.

FUNDAMENTAL SYSTEM THEORETIC CONCEPTS

Operators

- Define operators $q_h[\cdot]$ and $q_v[\cdot]$ as

$$q_h[\mathbf{x}](i,j) = \mathbf{x}(i+1,j) \quad \text{and} \quad q_v[\mathbf{x}](i,j) = \mathbf{x}(i,j+1).$$

- Propose operators $\delta_h[\cdot]$ and $\delta_v[\cdot]$ as

$$\delta_h[\mathbf{x}](i,j) = \frac{q_h[\mathbf{x}](i,j) - \mathbf{x}(i,j)}{\Delta_h};$$
$$\delta_v[\mathbf{x}](i,j) = \frac{q_v[\mathbf{x}](i,j) - \mathbf{x}(i,j)}{\Delta_v}.$$

Here, Δ_h and Δ_v are positive constants (that are the counterparts of sampling time).

q-Oerator Based Roesser Model

q-operator based Roesser model $\{A_q, B_q, C_q, D_q\}$ of a linear, shift-invariant, strictly causal, p-input, q-output 2-D discrete-time system:

$$\begin{bmatrix} q_h[\mathbf{x}^h](i,j) \\ q_v[\mathbf{x}^v](i,j) \end{bmatrix} = \begin{bmatrix} A_q^{(1)} & A_q^{(2)} \\ A_q^{(3)} & A_q^{(4)} \end{bmatrix} \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} + \begin{bmatrix} B_q^{(1)} \\ B_q^{(2)} \end{bmatrix} \mathbf{u}(i,j)$$

$$\doteq [A_q] \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} + [B_q]\mathbf{u}(i,j);$$

$$\mathbf{y}(i,j) = \begin{bmatrix} C_q^{(1)} & C_q^{(2)} \end{bmatrix} \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} + [D_q]\mathbf{u}(i,j)$$

$$\doteq [C_q] \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} + [D_q]\mathbf{u}(i,j).$$

δ-Operator Based Roesser Model

We propose the following δ-operator based Roesser model:

Intermediate equation

$$\begin{bmatrix} \delta_h[\mathbf{x}^h](i,j) \\ \delta_v[\mathbf{x}^v](i,j) \end{bmatrix} = \begin{bmatrix} A_\delta^{(1)} & A_\delta^{(2)} \\ A_\delta^{(3)} & A_\delta^{(4)} \end{bmatrix} \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} + \begin{bmatrix} B_\delta^{(1)} \\ B_\delta^{(2)} \end{bmatrix} \mathbf{u}(i,j)$$

$$\doteq [A_\delta] \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} + [B_\delta]\mathbf{u}(i,j);$$

$$\mathbf{y}(i,j) = \begin{bmatrix} C_\delta^{(1)} & C_\delta^{(2)} \end{bmatrix} \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} + [D_\delta]\mathbf{u}(i,j)$$

$$\doteq [C_\delta] \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} + [D_\delta]\mathbf{u}(i,j).$$

Update equation

$$q_h[\mathbf{x}^h](i,j) = \mathbf{x}^h(i,j) + \Delta_h \cdot \delta_h[\mathbf{x}^h](i,j);$$
$$q_v[\mathbf{x}^v](i,j) = \mathbf{x}^v(i,j) + \Delta_v \cdot \delta_v[\mathbf{x}^v](i,j).$$

- $\{A_q, B_q, C_q, D_q\}$ and $\{A_\delta, B_\delta, C_\delta, D_\delta\}$ are related by

$$A_q = I + \tau A_\delta; \quad B_q = \tau B_\delta; \quad C_q = C_\delta; \quad D_q = D_\delta.$$

Here, $\tau = [\Delta_h I \oplus \Delta_v I]$.

Gramians

Analogous to the 1-D and 2-D q-operator cases, reachability and observability gramian are proposed as:

$$P \doteq \begin{bmatrix} P^{(1)} & P^{(2)} \\ P^{(3)} & P^{(4)} \end{bmatrix} = \frac{1}{(2\pi j)^2} \oint_{\mathcal{T}_\delta^2} FF^* \frac{dc_h}{1+\Delta_h c_h} \frac{dc_v}{1+\Delta_v c_v};$$

$$Q \doteq \begin{bmatrix} Q^{(1)} & Q^{(2)} \\ Q^{(3)} & Q^{(4)} \end{bmatrix} = \frac{1}{(2\pi j)^2} \oint_{\mathcal{T}_\delta^2} G^*G \frac{dc_h}{1+\Delta_h c_h} \frac{dc_v}{1+\Delta_v c_v}.$$

Here, $F(c_h, c_v) \doteq (I - A_\delta)^{-1} B_\delta$ and $G(c_h, c_v) \doteq C_\delta (I - A_\delta)^{-1}$. $\mathcal{T}_\delta^2$ denotes stability boundary.

Balanced Realizations

It is proposed to call $\{A_\delta, B_\delta, C_\delta, D_\delta\}$ *balanced* if

$$P^{(1)} = Q^{(1)} = \text{diag}\{\sigma_1^{(1)}, \ldots, \sigma_{n_h}^{(1)}\};$$
$$P^{(4)} = Q^{(4)} = \text{diag}\{\sigma_1^{(4)}, \ldots, \sigma_{n_v}^{(4)}\}.$$

Accomplishments

- Characteristic equation and transfer function, relationship with q-model, equivalent transformations, algorithm to check stability, etc., are developed.
- Computation of gramians is addressed. For separable systems, they are block diagonal and may be computed via solution of four Lyapunov equations.

COEFFICIENT SENSITIVITY

Coefficient sensitivity of proposed model is investigated via suitable differential sensitivity measures. Small perturbations are assumed.

FXP Case

Coefficient perturbation is approximately independent of its nominal value. Hence, define

$$M_{\mathrm{FXP}} \doteq \|S_{A_\delta}\|_1^2 + \frac{1}{p}\|S_{B_\delta}\|_2^2 + \frac{1}{q}\|S_{C_\delta}\|_2^2 + \frac{1}{pq}\|S_{D_\delta}\|_2^2.$$

Here, $S_{A_\delta} = \partial H_\delta / \partial A_\delta$, etc. H_δ is the transfer function.

Accomplishments

- Realizations that are bound optimal with respect to M_{FXP} are in fact balanced.
- When $\Delta_h < 1$ and $\Delta_v < 1$, that is, with fast 'sampling', balanced δ-model is better than its corresponding q-model.

FLP Case

Coefficient perturbation is approximately proportional to its nominal value. Hence, define

$$M_{\mathrm{FLP}} \doteq \|\tilde{S}_{A_\delta}\|_1^2 + \frac{1}{p}\|\tilde{S}_{B_\delta}\|_2^2 + \frac{1}{q}\|\tilde{S}_{C_\delta}\|_2^2 + \frac{1}{pq}\|\tilde{S}_{D_\delta}\|_2^2.$$

Here, $\tilde{S}_{A_\delta} = \sum\sum a_{ij_\delta} \partial H_\delta / \partial a_{ij_\delta}$, etc.

Accomplishments

- Realization that are bound optimal with respect to M_{FLP} are better than its corresponding q-model if

$$\|A_q - I\|_F^2 < \|A_q\|_F^2.$$

- High speed narrowband digital filters typically satisfy this requirement.

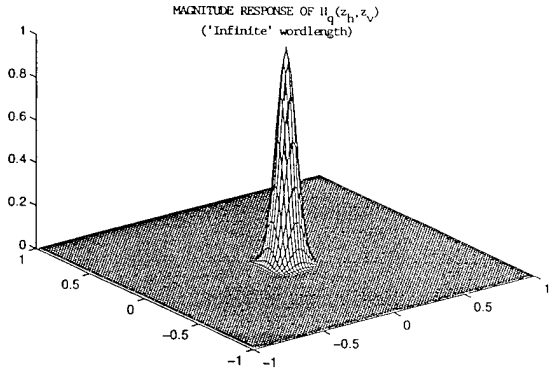
Example: Narrowband 5h-5v 2-D separable digital filter

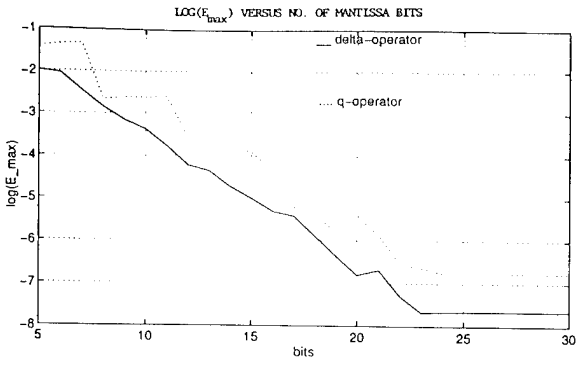
The corresponding q-Roesser model and transfer function are denoted as $\{A_q, B_q, C_q, D_q\}$ and $H_q(z_h, z_v)$, respectively.

- Let $\{\tilde{A}_q, \tilde{B}_q, \tilde{C}_q, \tilde{D}_q\}$ denote the corresponding balanced q-system. Under finite wordlength, let the transfer function be $\tilde{H}_q(z_h, z_v)$.
- Let $\{\tilde{A}_\delta, \tilde{B}_\delta, \tilde{C}_\delta, \tilde{D}_\delta\}$ denote the corresponding balanced δ-system. Under finite wordlength, let the transfer function be $\tilde{H}_\delta(c_h, c_v)$.

Simulation Data

- Mantissa length: $\tilde{H}_q$ and $\tilde{H}_\delta$ were implemented with different mantissa lengths (for the coefficients) to see the effects of coefficient sensitivity.
- Plot shows $\log[E_{\max}]$ versus mantissa length. Here, $\log[E_{\max}] = |\tilde{H}_q - H_q|$ (for the q-system) or $\log[E_{\max}] = |\tilde{H}_\delta - H_q|$ (for the δ-system). H_q is implemented with 'infinite' wordlength.
- Clearly, balanced δ-system performs better than the balanced q-system (which is 'optimal' with respect to the q-system counterpart of sensitivity measure M_{FXP})!





COMPARISON OF IMPLEMENTATIONS

	FXP CASE	FLP CASE
Quantization error bounds General system	δ-systems mostly superior	δ-systems better than or equal to q-systems
Quantization error bounds Digital equivalent of continuous-time system with short sampling time	δ-systems mostly superior	δ-systems superior
Limit cycles	δ-systems exhibit limit cycles	q- and δ-systems both exhibit limit cycles only in underflow
Dynamic range constraints Register overflow	δ-systems more likely to cause overflow	Unlikely in both q- and δ-systems
Coefficient sensitivity General system	δ-systems superior	δ-systems better than or equal to q-systems
Coefficient sensitivity Digital equivalent of continuous-time system with short sampling time	δ-systems superior	δ-systems superior
Hardware requirements Implementation of δ^{-1} requires additional sum and product	δ-systems require longer registers (for both coefficients and signals)	q- and δ-systems comparable

Appendix B: Other Papers/Presentations Where Grant #N00014-94-1-0454 Is Acknowledged

List of Papers/Presentations

Journal papers:

[B1] Premaratne, K., and Jury, E.I., "Discrete-time positive-real lemma revisited: The discrete-time counterpart of the Kalman-Yakubovitch lemma," *IEEE Trans. Circ. Syst.—I: Fund. Theo. Appl.,* **41**, pp. 747-750, Nov. 1994.

[B2] Premaratne, K., and Mansour, M., "Robust stability of time-variant discrete-time systems with bounded parameter perturbations," *IEEE Trans. Circ. Syst.—I: Fund. Theo. Appl.,* **42**, pp. 40-45, Jan. 1995.

[B3] Ekanayake, M.M., and Premaratne, K., "Two-channel IIR QMF banks with approximately linear-phase analysis and synthesis filters," *IEEE Trans. Sig. Proc.,* 1995, in review.

[B4] Corral, C.A., Lindquist, C.S., and Premaratne, K., "Minimax algorithm for optimum constrained block matrix filters," *IEEE Trans. Sig. Proc.,* 1995, in review.

simulations indicate more complex behavior than is exhibited by Type-I circuits. However, the existence of an extra port (port 1) in Type-II references may permit substitution of alternative devices or subcircuits to effect temperature- or V_{dd}-compensation. In addition, the voltage-following property may permit trimming of the initial operating point without effecting changes in the compensation network which stabilizes the RTD.

A new technique for generating well regulated reference currents and reference voltages is presented, applied here to MESFET circuits but applicable to any DFET technology. Simulation results suggest that supply rejection in these circuits may be comparable to the best present all-MESFET reference circuits.

ACKNOWLEDGMENT

The author is indebted to Professor James K. Roberge of the MIT EECS Department for several helpful discussions and sharing of technical insight, to Dr. Elliot Brown of MIT Lincoln Laboratory for sharing data on room temperature stable RTDs prior to publication, and to Ms. Lori Pressman of the MIT Technology Licensing Office for continued interest in this work.

REFERENCES

[1] C. Toumazou and D. G. Haigh, "Design and application of GaAs MESFET current mirror circuits," in *IEE Proc. Part G*, vol. 137, no. 2, pp. 101–108, Apr. 1990.
[2] R. Pflueger, "New RTD-bootstrapped current and voltage references I. Self-bootstrapped references, *IEEE Trans. Circuits and Syst. I*, vol. 41, pp. 740–743, Nov. 1994.
[3] R. Pflueger, "A bootstrap voltage reference based upon an N-type negative resistance device," *IEEE Trans. Instrumentation and Measurement*, vol. 42, pp. 719–725, June 1993.
[4] J. Y. Li and F. G. Weiss, "GaAs voltage reference generator," U.S. Patent 4 686 451, August 11, 1987.

Discrete-Time Positive-Real Lemma Revisited: The Discrete-Time Counterpart of the Kalman-Yakubovitch Lemma

Kamal Premaratne, and Eliahu I. Jury

Abstract—**In this paper, we present what can be considered the discrete-time counterpart of the concept of positive realness and the corresponding algebraic necessary and sufficient criteria that a discrete-time transfer function matrix must satisfy. This latter result may be thought of as the discrete-time counterpart of the Kalman-Yakubovitch Lemma and it is expected to find application in various areas of study. In particular, its use in proving the Jury-Lee criterion applicable in absolute stability studies of a certain class of discrete-time nonlinear systems is shown.**

I. INTRODUCTION

The Kalman-Yakubovitch (KY) Lemma [1-2] applicable in continuous-time (CT) systems and its discrete-time (DT) version [3] have found application in various areas of study, such as, network synthesis, spectral factorization, nonlinear system stability, etc (see [4], and references therein).

Manuscript received January 5, 1994; revised May 8, 1994. This work was partially supported by the Office of Naval Research (ONR) through the grant N00014-94-1-0454.

The authors are with the Department of Electrical and Computer Engineering, P.O. Box 248294, University of Miami, Coral Gables, FL 33124 USA.

IEEE Log Number 9405610.

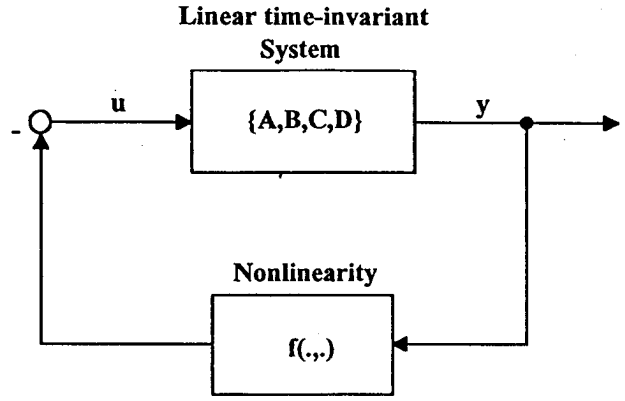


Fig. 1.

However, this DT version may not be used to prove the DT counterpart to the Popov criterion applicable in absolute stability studies of CT systems, that is, the Jury-Lee criterion [5]. Hence, it cannot be considered a true DT *counterpart* of the KY Lemma. In fact, according to terminology established in [6], the DT version in [3] is the DT *analog* of the KY Lemma.

The results in this paper presents what can be considered the DT counterpart of positive-realness and the corresponding algebraic necessary and sufficient conditions. This constitutes the solution of an outstanding research problem that has attracted considerable attention [7-9]. The result presented may be used to prove the Jury-Lee criterion, and hence, it can be considered the DT *counterpart* of the KY Lemma.

II. PRELIMINARIES

2.1. Nomenclature

Real and complex numbers are denoted by $\Re$ and $\Im$, respectively. The sets of matrices of size $p \times q$ over $\Re$ and $\Im$ are $\Re^{p\times q}$ and $\Im^{p\times q}$, respectively. The set of rational polynomials over $\Re$ is denoted by $\Re(z)_{p\times q}$. Complex conjugate of $z \in \Im$ is z^*.

P^* and P^T are the conjugate transpose and transpose of $P \in \Re^{p\times q}$, respectively. $P \in \Re^{p\times p}$ being positive definite is denoted by $P > 0$. If $P \in \Im^{p\times q}$, $[\![P]\!] \doteq P + P^*$; if $P \in \Re^{p\times q}$, $[\![P]\!] \doteq P + P^T$. $G^*(z)$ is the complex conjugate transpose of $G(z) \in \Re(z)_{p\times q}$. Hence, $G^*(z) = G(z^*)^T$. Normal rank of $G(z) \in \Re(z)_{p\times p}$ is $\mathrm{rank}[G(z)]$.

The sets $\{z \in \Im : |z| = 1\}$ and $\{z \in \Im : |z| > 1\}$ are denoted by $\mathcal{T}_q$ and $\mathrm{ext}[\mathcal{T}_q]$, respectively.

2.2. Absolute Stability

Often, nonlinear systems arising in practice may be represented in the form of a linear time-invariant system with a possibly time-variant memoryless nonlinearity in the feedback path [10-12]. See Fig. 1. Consider the case when this linear part is a DT system possessing the minimal state-space realization $\{A, B, C, D\}$ where

$$q[\mathbf{x}](i) = A\mathbf{x}(i) + B\mathbf{u}(i); \quad \mathbf{y}(i) = C\mathbf{x}(i) + D\mathbf{u}(i), \qquad (2.1)$$

and

$$\mathbf{u}(i) = -\mathbf{f}(i, \mathbf{y}(i)). \tag{2.2}$$

Here, $\mathbf{u}, \mathbf{y} \in \Re^p$, $\mathbf{x} \in \Re^n$, and $q[\mathbf{x}](i) \doteq \mathbf{x}(i+1)$. Let the nonlinearity $\mathbf{f} : [0, \infty) \times \Re^p \to \Re^p : t \times \mathbf{y} \to \mathbf{f}(t, \mathbf{y})$ be possibly time-variant, memoryless, piecewise continuous in t, and locally Lipschitz in $\mathbf{y}$. The transfer function of the DT linear part is

$$G(z) = C\Gamma_A(z)^{-1}B + D \in \Re(z)_{p\times p} \quad \text{where} \quad \Gamma_A(z) \doteq (zI - A). \tag{2.3}$$

In stability studies of such systems, $\mathbf{f}$ is typically taken to belong to the following class of functions:

Definition 2.1 [12]: Let $\mathbf{f} : [0, \infty) \times \Re^p \to \Re^p : t \times \mathbf{y} \to \mathbf{f}(t, \mathbf{y})$ be a memoryless nonlinearity, and $\underline{K}$ and $\overline{K}$ be symmetric matrices in $\Re^{p\times p}$ with $\overline{K} - \underline{K} > 0$. Then, $\mathbf{f}$ is said to belong to the *sector* $[\underline{K}, \overline{K}]$, or $\mathbf{f} \in [\underline{K}, \overline{K}]$, if (1) $\mathbf{f}(t, \mathbf{0}) = \mathbf{0}$, $\forall t \in [0, \infty)$, and (2) $[\mathbf{f}(t, \mathbf{y}(t)) - \underline{K}\mathbf{y}]^T[\mathbf{f}(t, \mathbf{y}(t)) - \overline{K}\mathbf{y}] \leq 0$, $\forall t \in [0, \infty)$, $\forall \mathbf{y} \in \Re^p$.

Given the configuration in Fig. 1 with $\mathbf{f} \in [\underline{K}, \overline{K}]$, what conditions must be imposed on the frequency response of $G(z)$ to ensure global uniform asymptotic stability? This is generally referred to as the *Lur'e problem*[1]. When these conditions are satisfied at the origin, the configuration in Fig. 1 is said to be *absolutely stable.*

In solving the Lur'e problem, two types of Lyapunov function candidates are generally used. The *quadratic form* $V(\mathbf{x}) = \mathbf{x}^T P\mathbf{x}$, where $P = P^T > 0$ and the nonlinearity is possibly time-variant, yields the *Tsypkin criterion* [10], [13]. This is the analog of the *circle criterion* applicable in CT systems [12] which may be proven using the following important result [1-2]:

Lemma 2.1. Kalman-Yakubovitch Lemma: Consider the minimal realization $\{A, B, C, D\}$ of a stable CT transfer function $G(s) \in \Re(s)_{p\times p}$. Then, $G(s)$ is strictly positive-real iff, for some $\epsilon > 0$, $\exists\, P = P^T > 0$ and matrices W and L such that (1) $A^TP + PA = -L^TL - \epsilon P$; (2) $B^TP = C - W^TL$; and (3) $W^TW = D + D^T$.

The *DT analog* of the above may be used to prove the Tsypkin criterion [3]:

Lemma 2.2. DT Analog of the Kalman-Yakubovitch Lemma: Consider the minimal realization $\{A, B, C, D\}$ of a stable DT transfer function $G(z) \in \Re(z)_{p\times p}$. Then, $G(z)$ is strictly positive-real iff $\exists\, P = P^T > 0$ and matrices W and L such that (1) $A^TPA - P = -L^TL$; (2) $B^TPA = C - W^TL$; and (3) $W^TW = D + D^T - B^TPB$.

The above criteria, in some cases, tend to provide conditions that are conservative [5], [10-11]. Hence, an additional restriction on the slope of the nonlinearity, together with the *Lur'e form* $V(\mathbf{x}) = \mathbf{x}^TP\mathbf{x} + \beta \int_{\zeta=0}^{\mathbf{y}} \mathbf{f}^T(\zeta)K\, d\zeta$, where $P = P^T > 0$, $\beta \in [0, \infty)$, $K = \overline{K} - \underline{K} > 0$, and the nonlinearity is time-invariant, has been utilized [5], [11], [14-15]. This result is generally referred to as the *Jury-Lee criterion* [7], [16]. Simpler proofs of this, for special cases, appear in [14], [16-17].

The Jury-Lee criterion, as sampling frequency increases, yields the Popov criterion applicable in the CT case [5], [13] which may also be proven using the KY Lemma [1-2]. Hence, the Jury-Lee criterion may be considered the DT counterpart of the Popov criterion [6]. Since the DT *analog* of the KY Lemma (Lemma 2.2) cannot be used to prove the Jury-Lee criterion, is there a DT *counterpart* to the KY Lemma that will serve the same purpose? Several previous attempts addressing this [7-8] contained discrepancies that were later pointed out [9].

In the following, what can be considered the DT counterpart of positive-realness is presented. Next, a related result that facilitates an easier and more direct proof of the Jury-Lee criterion, with possibly multivariable and nonisolated nonlinearities, is given. Hence, it may be considered a 'true' DT counterpart to the KY Lemma thus providing a solution to an outstanding problem.

[1]The Lur'e problem was originally posed for CT systems. What is being posed here is its DT version.

III. Discrete-Time Counterpart of Positive-Realness

What may be considered the DT counterpart of positive-realness is defined as follows:

Definition 3.1: Consider a proper square matrix $G(z) \in \Re(z)_{p\times p}$ with a minimal realization $\{A, B, C, D\}$ as in (2.1). $G(z)$ is said to be *discrete-time positive-real* iff (1) $G(z)$ is analytic in ext$[\mathcal{T}_q]$; and (2) $[\![(I+(z-1)V)(G(z)-D)+(I-V)D]\!] - |(z-1)(G(z)-D)|^2 \geq 0$ in ext$[\mathcal{T}_q]$. Here, $V \in \Re^{p\times p}$ is an arbitrary matrix.

As in [3], the above may be restated as follows:

Lemma 3.1: Let $G(z)$ be as in Definition 3.1. It is DT positive-real iff (1) $G(z)$ is analytic in ext$[\mathcal{T}_q]$; (2) on $\mathcal{T}_q$, $G(z)$ has only simple poles, and at these locations, the corresponding residue matrix is Hermitian positive semidefinite; and (3) $[\![(I + (z-1)V)(G(z) - D) + (I-V)D]\!] - |(z-1)(G(z)-D)|^2 \geq 0$ on $\mathcal{T}_q$ whenever $G(z)$ exists. Here, $V \in \Re^{p\times p}$ is an arbitrary matrix.

Equations (3.1-2) and Theorem 3.2 will be used in the ensuing discussion:

$$P - A^TPA = \Gamma_A^*(z)P\Gamma_A(z) + [\![A^TP\Gamma_A(z)]\!], \ \forall z \in \mathcal{T}_q \tag{3.1}$$

and

$$CA\Gamma_A(z)^{-1}B = z(G(z) - D) - CB. \tag{3.2}$$

Theorem 3.2. Spectral Factorization Theorem: Consider a proper square matrix $V(z) \in \Re(z)_{p\times p}$. Suppose $V(z) = V^*(z)$ and $V(z) > 0$, $\forall z \in \mathcal{T}_q$. Then, $\exists$ a proper stable $T(z) \in \Re(z)_{p\times p}$ such that $V(z) = T^*(z)T(z)$. Moreover, rank$[T(z)] = p$, $\forall z \in \mathcal{T}_q$.

For convenience, from now on, we assume that A is stable. Algebraic necessary and sufficient criteria to satisfy item (3) in Lemma 3.1 are now given as follows:

Theorem 3.3: Let $G(z)$ be as in Definition 3.1 with A being stable. Then, the following two conditions are equivalent:

(A.) *Frequency domain condition:* $G(z)$ is DT positive-real, that is, $[\![(I + (z-1)V)(G(z) - D) + (I - V)D]\!] - |(z-1)(G(z)-D)|^2 \geq 0$, $\forall z \in \mathcal{T}_q$. Here, $V \in \Re^{p\times p}$ is arbitrary.

(B.) *DT counterpart of the KY Lemma:* There exists $P = P^T > 0$ and matrices W and Q such that

$$-Q^TQ = A^TPA - P + (A - I)^TC^TC(A - I); \tag{B1}$$

$$-W^TQ = B^TPA!-!C!-!VC(A!-!I)!+!B^TC^TC(A!-!I); \tag{B2}$$

$$-W^TW = B^TPB + B^TC^TCB - [\![VCB + (I - V)D]\!]. \tag{B3}$$

Proof: (B) implies (A): Premultiplying by $B^T\Gamma_A^*(z)^{-1}$ and postmultiplying by $\Gamma_A^{-1}B$, and then using (3.1), (B1) yields

$$\begin{aligned} B^TPB + [\![B^TPA\Gamma_A^{-1}B]\!] &= B^T\Gamma_A^{*^{-1}}Q^TQ\Gamma_A^{-1}B + B^T\Gamma_A^{*^{-1}}(A - I)^T \\ &\cdot C^TC(A - I)\Gamma_A^{-1}B. \end{aligned} \tag{3.3}$$

Note that, using (3.2), we have

$$\begin{aligned} B^T\Gamma_A^{*^{-1}}(A - I)^TC^TC(A - I)\Gamma_A^{-1}B &= |(z-1)(G(z) - D)|^2 \\ &- B^TC^TCB - [\![B^TC^TC(A - I)\Gamma_A^{-1}B]\!]. \end{aligned} \tag{3.4}$$

Substituting (3.4) in (3.3) yields (3.5), shown at the bottom of the next page. Using B^TPA and B^TPB from (B2) and (B3), respectively, and (3.2), we get (A).

(A) implies (B): Let

$$H_1(z) = (I + (z-1)V)(G(z) - D) + (I - V)D \\ = (C + VC(A-I))\Gamma_A^{-1}B + VCB + (I - V)D \quad (3.6)$$

and $T_2(z) = (z-1)(G(z) - D) = C(A-I)\Gamma_A^{-1}B + CB$, where we have used (3.2). Hence

$$T_2^*(z)T_2(z) = |(z-1)(G(z) - D)|^2 \\ = B^T\Gamma_A^{*-1}(A-I)^TC^TC(A-I)\Gamma_A^{-1}B + B^TC^TCB \\ + [\![B^TC^TC(A-I)\Gamma_A^{-1}B]\!]. \quad (3.7)$$

Noting that $\{C(A-I), A\}$ is observable (since $\{C, A\}$ is), $\exists R = R^T > 0$ such that

$$(A-I)^TC^TC(A-I) = R - A^TRA = \Gamma_A^*R\Gamma_A + [\![A^TR\Gamma_A]\!] \quad (3.8)$$

where we have used (3.1). Substituting (3.8) in (3.7) yields

$$T_2^*(z)T_2(z) = B^T(R + C^TC)B \\ + [\![B^T(RA + C^TC(A-I))\Gamma_A^{-1}B]\!]. \quad (3.9)$$

Then, $[\![H_2(z)]\!] = T_2^*(z)T_2(z)$, where

$$H_2(z) = B^T(RA + C^TC(A-I))\Gamma_A^{-1}B + \frac{1}{2}B^T(R + C^TC)B. \quad (3.10)$$

Now, with $H(z) \doteq H_1(z) - H_2(z)$, $[\![H(z)]\!]$ being the left hand side of (A), we have

$$[\![H(z)]\!] \geq 0, \ \forall z \in \mathcal{T}_q. \quad (3.11)$$

Letting

$$V(z) = [\![H(z)]\!] \Longrightarrow V(z) = V^*(z); \ V(z) \geq 0, \ \forall z \in \mathcal{T}_q. \quad (3.12)$$

Thus, from Theorem 3.2, $\exists$ a proper stable $T(z) \in \Re(z)_{p\times p}$ such that

$$V(z) = T^*(z)T(z). \quad (3.13)$$

Let $\{F, G, K, L\}$ be a minimal realization of $T(z)$, that is, $T(z) = K\Gamma_F^{-1}G + L$. Hence

$$T^*(z)T(z) = G^T\Gamma_F^{*-1}K^TK\Gamma_F^{-1}G + L^TL + [\![L^TK\Gamma_F^{-1}G]\!]. \quad (3.14)$$

Since $\{K, F\}$ is observable, $\exists S = S^T > 0$ such that

$$K^TK = S - F^TSF = \Gamma_F^*S\Gamma_F + [\![F^TS\Gamma_F]\!]. \quad (3.15)$$

Therefore, substituting (3.15) in (3.14) yields

$$T^*(z)T(z) = (G^TSG + L^TL) + [\![(G^TSF + L^TK)\Gamma_F^{-1}G]\!]. \quad (3.16)$$

From (3.11-13), $T^*(z)T(z) = V(z) = [\![H(z)]\!] = [\![H_1(z) - H_2(z)]\!]$. Hence (3.17), which is shown at the bottom of this page. Compare (3.16) and (3.17). Since F and A are stable, match the stable parts to yield

$$(G^TSF + L^TK)\Gamma_F^{-1}G = \\ \Big[C + VC(A-I) - B^T(RA + C^TC(A-I))\Big]\Gamma_A^{-1}B. \quad (3.18)$$

Hence, $\exists$ a nonsingular matrix $M \in \Re^{n\times n}$ such that

$$M^{-1}FM = A; \quad M^{-1}G = B; \\ (G^TSF + L^TK)M = C + VC(A-I) - B^T(RA + C^TC(A-I)). \quad (3.19)$$

Now, define $P = M^TSM + R$, $Q = KM$, and $W = L$. Premultiplying by M^T and postmultiplying by M, (3.15) yields (B1) when the first equation in (3.19) and (3.8) are used. Last equation in (3.19) yields (B2). Compare the constant terms in (3.16-17) to get (B3). □

Note: The presence of V to be manipulated as an additional parameter will be useful in the next section when the Jury-Lee criterion is proven.

IV. Jury-Lee Criterion

Jury-Lee criterion may be used for absolute stability studies of multivariable DT nonlinear systems [15]. Corresponding sufficient conditions, as obtained in [16-17], are now derived using Theorem 3.3. In [17], the nonlinear system in (2.1) and (2.2) with $D = 0$ is considered. Let the nonlinearity possess the following properties: For all $i = 1, 2, \ldots, p$,

(1) $f_i(0) = 0$; (2) $0 < y_if_i(y_i) < k_iy_i^2$; and (3) $-\underline{\hat{k}}_i < \frac{df_i(y_i)}{dy_i} < \overline{\hat{k}}_i$. Here, $\underline{\hat{k}}_i \geq 0$, $\overline{\hat{k}}_i \geq k_i$, $\forall i = 1, 2, \ldots, p$. In [17], it is shown that, existence of $P = P^T > 0$ and matrices Q and R such that

$$A^TPA - P + (A-I)^TC^TM^TMC(A-I) = -Q^TQ; \\ B^TPA - NC - SC(A-I) + B^TC^TM^TMC(A-I) = -R^TQ; \\ B^TPB + B^TC^TM^TMCB - 2NK^{-1} - [\![SCB]\!] = -R^TR \quad (4.1)$$

is sufficient for absolute stability. Here, M is a certain diagonal matrix with positive elements associated with the slope condition, $N = N^T > 0$, and S is a diagonal matrix. Also, $K = \text{diag}\{k_1, \ldots, k_p\}$.

$$B^TPB + [\![B^TPA\Gamma_A^{-1}B]\!] = \left|Q\Gamma_A^{-1}B\right|^2 + |(z-1)(G(z) - D)|^2 - B^TC^TCB \\ - [\![B^TC^TC(A-I)\Gamma_A^{-1}B]\!] \\ = \left|Q\Gamma_A^{-1}B + W\right|^2 + |(z-1)(G(z) - D)|^2 - B^TC^TCB \\ - W^TW - [\![W^TQ + B^TC^TC(A-I)\Gamma_A^{-1}B]\!]. \quad (3.5)$$

$$T^*(z)T(z) = [\![\Big[C + VC(A-I) - B^T(RA + C^TC(A-I))\Big]\Gamma_A^{-1}B + VCB + (I-V)D]\!] \\ - B^T(R + C^TC)B. \quad (3.17)$$

Substitute the following in (B1), (B2), and (B3) in Theorem 3.3: $A \rightarrow A$; $B \rightarrow BN^{T^{-1}}M^T$; $C \rightarrow MC$; $[(I - V)D] \rightarrow 2MK^{-1}N^{-1}M^T$. Now, (A) yields the following frequency domain condition which is identical to that in [17]:

$$NK^{-1} + [\frac{1}{2}[N + (z-1)S]G(z)] - \frac{1}{2}|z-1|^2 \cdot G^*(z)M^T MG(z) \geq 0, \ \forall z \in \mathcal{T}_q. \tag{4.2}$$

V. Conclusion and Final Remarks

What may be considered the DT counterpart of positive-realness and the corresponding algebraic necessary and sufficient conditions (Theorem 3.3) have been presented. This latter result facilitates the proof of Jury-Lee criterion and can be thought of as the DT counterpart of the KY Lemma, thus successfully addressing an outstanding research problem [7-9]. It is also expected to find use in generalizing the Jury-Lee criterion and in various other areas of study, such as, network synthesis, spectral factorization, etc.

References

[1] V. A. Yakubovitch, "The solution of certain matrix inequalities in automatic control theory," *Doklady Akademii Nauk SSSR,* vol. 143, pp. 1304–1307, 1962.

[2] R. E. Kalman, "Lyapunov functions for the problem of Lur'e in automatic control," in *Proc. Nat. Acad. Sci.,* vol. 49, pp. 201–205, 1963.

[3] L. Hitz and B.D.O. Anderson, "Discrete positive-real functions and their application to system stability," in *Proc. IEE,* vol. 116, pp. 153–155, Jan. 1969.

[4] B.D.O. Anderson, K. L. Hitz, and N. D. Diem, "Recursive algorithm for spectral factorization," *IEEE Trans. Circuits and Syst.,* vol. CAS-21, pp. 742–750, Nov. 1974.

[5] E. I. Jury and B. W. Lee, "On the stability of a certain class of nonlinear sampled-data systems," *IEEE Trans. Automat. Cont.,* vol. AC-9, pp. 51–61, Jan. 1964.

[6] E. I. Jury and M. Mansour, "On the terminology relationship between continuous and discrete system criteria," *Proc. IEEE,* vol. 73, p. 844, Apr. 1985.

[7] M.K.P. Mishra and A. K. Mahalanabis, "On the stability of discrete nonlinear feedback systems with state-dependant noise," *Intl. J. Syst. Sci.,* vol. 6, pp. 479–490, 1975.

[8] N. O'Reilly and P. C. Byrne, "Frequency-domain condition for the existence of a Lyapunov functional for the problem of Lur'e," *IEEE Trans. Auto. Control,* vol. AC-25, pp. 555–557, June 1980.

[9] V. Singh, "Extended MKY lemma—What shall it be?" *IEEE Trans. Automat. Cont.,* vol. AC-28, pp. 627–628, May 1983.

[10] S. Kodama, "Stability of nonlinear sampled-data systems," *IRE Trans. Automat. Cont.,* vol. AC-7, pp. 15–23, Jan. 1962.

[11] J. B. Pearson and J. E. Gibson, "On the asymptotic stability of a class of saturating sampled-data systems," *IEEE Trans. Appl. Industry,* vol. AI-83, pp. 81–86, Mar. 1964.

[12] H. K. Khalil, *Nonlinear Systems.* New York: Macmillan, 1992.

[13] Y. Z. Tsypkin, "The absolute stability of large-scale, nonlinear, sampled-data systems," *Doklady Akademii Nauk SSSR,* vol. 145, pp. 52–55, July 1962.

[14] G. P. Szegö, "On the absolute stability of sampled-data control systems," in *Proc. Nat. Acad. Sci.,* vol. 50, pp. 558–560, 1963.

[15] E. I. Jury and B. W. Lee, "The absolute stability of systems with many nonlinearities," *Avtomatika i Telemekhanika,* vol. 16, pp. 945–965, Jun. 1965.

[16] V. Singh, "Lyapunov-based proof of Jury-Lee's criterion: Some appraisals," *IEEE Trans. Circuits and Syst.,* vol. CAS-32, pp. 396–398, Apr. 1985.

[17] T. N. Sharma and V. Singh, "On the absolute stability of multivariable discrete-time nonlinear systems," *IEEE Trans. Automat. Cont.,* vol. AC-26, pp. 585–586, Apr. 1981.

Impact of Distributed Gate Resistance on the Performance of MOS Devices

Behzad Razavi, Ran-Hong Yan, and Kwing F. Lee

***Abstract*—This paper describes the impact of gate resistance on cut-off frequency (f_T), maximum frequency of oscillation (f_{max}), thermal noise, and time response of wide MOS devices with deep submicron channel lengths. The value of f_T is proven to be independent of gate resistance even for distributed structures. An exact relation for f_{max} is derived and it is shown that, to predict f_{max}, thermal noise, and time response, the distributed gate resistance can be divided by a factor of 3 and lumped into a single resistor in series with the gate terminal.**

I. Introduction

The remarkable improvement in the performance of CMOS circuits as a result of scaling has motivated extensive research on deep submicron MOS devices [1], [2]. While short channel effects such as velocity saturation and threshold voltage variation become significant for dimensions below approximately 2 μm, other nonidealities manifest themselves for only very small channel lengths. In particular, the gate resistance of a short-channel device can substantially affect its performance if the transistor width is increased to attain high current drive or large transconductance. This effect becomes especially noticeable in line drivers and output buffers used in digital systems and low-noise, high-gain amplifiers employed in analog applications, all of which typically require wide MOSFETs.

Even though the overall gate resistance can be lowered through silicidation or the use of multiple gates, these remedies have certain limitations. For example, the thickness of gate silicide must scale with channel length, thereby yielding a higher sheet resistivity for shorter devices [1]. Also, increasing the number of gates (to allow narrower devices for a given total width) tends to increase the source or drain junction capacitance and degrade circuit density.

This paper describes the impact of distributed gate resistance on four aspects of the performance of deep submicron devices: cut-off frequency (f_T), maximum frequency of oscillation (f_{max}), input-referred thermal noise, and time response. The primary goal is to quantify this impact with relatively simple expressions, thus allowing technologists and circuit designers to easily determine the maximum gate resistance that can be tolerated in a given application. The analyses are performed for an NMOS transistor whose gate is contacted only at one end, but the results can be readily applied to all field effect devices and structures with multiple gate contacts as well.

The next section of the paper analyzes the effect of gate resistance on f_T. Sections III to V, respectively, formulate the dependence of f_{max}, thermal noise, and time response on the gate resistance. Section VI summarizes the results.

II. Cut-off Frequency

Defined as the frequency at which the short-circuit small-signal current gain of a transistor drops to unity, f_T is a measure of the speed of the intrinsic device excluding its junction parasitics. If the gate resistance is modelled as a lumped resistor in series with the

Manuscript received June 14, 1993; revised January 24, 1994. This paper was recommended by Associate Editor David Haigh.

The authors are with AT&T Bell Laboratories, Room 4E312, 101 Crawfords Corner Road, Holmdel, NJ 07733 USA.

IEEE Log Number 9405609.

[3] M. L. Liou, "Exact analysis of linear circuits containing periodically operated switches with applications," *IEEE Trans. Circuit Theory*, vol. 19, pp. 146–154, Mar. 1972.
[4] T. Storm and S. Signell, "Analysis of periodically switched linear circuits," *IEEE Trans. Circuits Syst.*, vol. 24, pp. 531–540, Oct. 1977.
[5] J. C. Candy and G. C. Temes, Ed., *Oversampling Delta-Sigma Data Converters*. New York: IEEE Press, 1992.
[6] J. H. Fischer, "Noise sources and calculation techniques for switched capacitor filters," *IEEE J. Solid-State Circuits*, vol. 17, pp. 742–752, Aug. 1972.
[7] C. A. Desoer and E. S. Kuh, *Basic Circuit Theory*. New York: McGraw Hill, 1969.
[8] G. Dahlquist and A. Bjorck, *Numerical Methods*. Englewood Cliffs, NJ: Prentice-Hall, 1974.
[9] J. Vlach and K. Singhal, *Computer Methods for Circuit Analysis and Design*, 2nd ed. New York: Van Nostrand Reinhold, 1993.
[10] J. A. Nossek and G. C. Temes, "Switched capacitor filter design using bilinear element modeling," *IEEE Trans. Circuits Syst.*, vol. 27, pp. 481–491, June 1980.

Robust Stability of Time-Variant Discrete-Time Systems with Bounded Parameter Perturbations

Kamal Premaratne and Mohamed Mansour

***Abstract*— In this paper, global asymptotic stability of linear, time-variant, finite dimensional, zero input difference equations is investigated. We propose a technique that may be utilized to obtain regions of asymptotic stability in the coefficient space that incorporate information regarding the maximum rate of change of system parameters. Use of different matrix norms provide different "shapes" for the maximum allowable coefficient perturbations.**

NOMENCLATURE

$\Re$, $\aleph$ Real and integer numbers.
$\aleph_+$, $\aleph_+^0$ Positive and non-negative integers.
$\Re^k$, $\Re^{k\times\ell}$ Vectors of length k over $\Re$, Matrices of size $k\times\ell$ over $\Re$.
$\mathbf{0}_{k\times\ell}$, $\mathbf{I}_{k\times\ell}$ Null matrix and identitiy matrix of size $k\times\ell$.

Given quantity $[\cdot]$ of a TV system, the analogous quantity of the corresponding TI system is denoted by $[\hat{\cdot}]$. For example,

$\mathbf{a}(n)\in\Re^m$ Coefficient vector $[a_1(n),\ldots,a_m(n)]^T$ of a TV difference equation of order m.
$\hat{\mathbf{a}}\in\Re^m$ Coefficient vector of the corresponding TI difference equation of order m, that is, $[a_1,\ldots,a_m]^T$.
$A(n)\in\Re^{m\times m}$ Companion form corresponding to $\mathbf{a}(n)$ (see (3.3)).
$\lambda_i[\hat{A}]$ Eigenvalues of $\hat{A}\in\Re^{m\times m}$.
$\|A\|_p$ Induced p-norm of $A\in\Re^{m\times m}$, that is, $\sup_{\mathbf{x}\neq\mathbf{0}}\frac{\|A\mathbf{x}\|_p}{\|\mathbf{x}\|_p}$. Here, $p\in\{\aleph_+\cup\infty\}$.
$\Delta_i(n)\in\Re$ Perturbation on the coefficient $a_i(n)\in\Re$ at time instant n.
$\Delta(n)\in\Re^{m\times m}$ Perturbation matrix on system matrix $A(n)\in\Re^{m\times m}$ at time instant n (see (3.17)).
$P_n^{(j+1)}$ Product of $j+1$ consecutive system matrices that are TV, that is, $\prod_{i=0}^{j}A(n+i)=A(n+j)\cdot A(n+j-1)\cdots A(n)\cdot A(n+j-1)\cdots A(n)$ (see (3.20)).

Manuscript received March 23, 1994; revised, July 26, 1994. This work was supported in part by the Office of Naval Research (ONR), grant N00014-94-1-0454. This paper was recommended by Associate Editor Guanrong Chen.

K. Premaratne is with the Department of Electrical and Computer Engineering, University of Miami, Coral Gables, FL 33124 USA.
M. Mansour was with the University of Miami, Coral Gables. He is now with the Automatic Control Laboratory, Swiss Federal Institute of Technology (ETH), CH-8092 Zürich, Switzerland.

IEEE Log Number 9407838.

II. INTRODUCTION

Parameter uncertainties are inherent in system models utilized for analysis and design. The explosion of recent research activity in related areas, in particular, in the area of robust stability, is mainly due to the seminal work of Kharitonov [1]. Since this result, robust stability of time-invariant (TI) systems with uncertain parameters have received considerable attention (see [2-3], and references therein).

Many important results regarding robust stability of time-variant (TV) systems with uncertain parameters are also available. Some earlier results appear in [4-6], and references therein; newer results are constantly being introduced (see [7-11], and references therein). Such systems find application in various branches in signal processing and control, such as, adaptive signal processing, finite wordlength implementation of digital filters [12], and design of reconfigurable systems [13].

For a TV system represented in its difference equation formulation, the work in [10] provides a region in the coefficient space wherein the coefficients may vary while maintaining global asymptotic stability (GAS). For a TV system represented in its state-space (SS) formulation, the work in [11] provides a necessary and sufficient condition for robust GAS. However, in these work, no restriction has been imposed on its maximum rate of change whereas, in most practical situations, such a restriction is typically inherent. An important outstanding research problem is to incorporate such information and obtain a region (or, regions) in the coefficient space where GAS of a TV system is guaranteed [2, Open Problem #9].

The work below attempts to address the above problem. The results presented, as they stand, can be computationally demanding. For second-order systems at least, it is quite conveniently applicable. The authors hope that this work may serve as an impetus for further improvements. The paper is organized as follows: Section II formulates the problem where, for the readers' convenience, we follow the same notation as in [10]. A different but enlightening proof of the main result in [10] is provided. Section III contains the main results. A procedure that can generate a region in the coefficient space that guarantees asymptotic stability (AS) is described. Section IV is provides an example. Section V contains concluding remarks.

III. PROBLEM FORMULATION

Consider the following linear, possibly time-variant (TV), finite dimensional, zero input, difference equation of order m:

$$y(n)=\sum_{i=1}^{m}a_i(n)y(n-i)=\mathbf{a}(n)^T\mathbf{y}(n-1). \tag{2.1}$$

Definition 2.1: The TV system in (2.1) is said to be *asymptotically stable in* Ω if, for any $\mathbf{y}(0)=[y(-1),\ldots,y(-m)]^T$, $\lim_{n\to\infty}\mathbf{y}(n)=0$ whenever $\mathbf{a}(n)\in\Omega$, $\forall n$.

Remarks:

1) Due to the fact that the system under consideration is linear, AS, as defined above, is equivalent to the notion of GAS [14, ch. 3].
2) In investigating such regions of AS, one may distinguish between TI and TV regions [10]. In this paper, our interest lies in obtaining only TI regions of AS.

Now, consider $\Omega \in \Re^m$. The problem of determining the "largest" such region so that, whenever $\mathbf{a}(n) \in \Omega$, $\forall n$, AS of (2.1) is guaranteed has been addressed in [10]. A relevant result is

Theorem II.1. [10]: Let $\Omega = \{\mathbf{a}(n) \in \Re^m : \sum_{i=1}^m |a_i(n)| \leq \gamma < 1, \forall n\}$. Then, whenever $\mathbf{a}(n) \in \Omega$, $\forall n$, the TV system in (2.1) is AS. Moreover, such a TI region for AS may not contain any point $\mathbf{a}(n)$ that satisfy $\sum_{i=1}^m |a_i(n)| = 1 + \epsilon$, for any $\epsilon > 0$.

Remark: The latter part of Theorem II.1 indicates that the region Ω is in fact the largest hyperdiamond region with the origin as its center. In this sense, the region Ω is in fact not too conservative.

Therefore, with no restriction imposed on the amount of perturbation allowable on each coefficient $a_i(n)$ *at each time instant,* as long as $\mathbf{a}(n) \in \Omega$, $\forall n$, AS of the system in (2.1) is guaranteed. However, in practice, the rate of change of each coefficient $a_i(n)$, is restricted. Can we incorporate this additional restriction into the above result? Intuitively, the region obtained thus must be larger than that indicated in Theorem II.1. It is this problem (see also [2], Open Problem #9) that we attempt to address.

IV. Main Results

By expressing the system in (2.1) in its SS formulation, we now prove Theorem II.1 utilizing norm arguments. Such a proof has the following advantages: a) It exposes a certain interesting norm property of companion matrices (see Lemma III.1); b) then, proof of Theorem II.1 follows quite easily; c) SS formulation is an ideal tool for the problem at hand; and d) it provides the possibility of using different norms thus yielding different regions of AS.

Clearly, utilizing the state variables

$$\begin{aligned} x_m(n) &= y(n); \\ x_{m-1}(n) &= x_m(n-1) = y(n-1); \\ &\vdots \\ x_1(n) &= x_2(n-1) = y(n-m+1), \end{aligned} \tag{3.1}$$

the system in (2.1) may be expressed as

$$\mathbf{x}(n) = A(n) \cdot \mathbf{x}(n-1); \qquad y(n) = C \cdot \mathbf{x}(n), \tag{3.2}$$

where $\mathbf{x}(n) = [x_1(n), \ldots, x_m(n)]^T$, and

$$A(n) = \begin{bmatrix} 0 & 1 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \\ a_m(n) & a_{m-1}(n) & \cdots & a_2(n) & a_1(n) \end{bmatrix} \in \Re^{m \times m};$$

$$C = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}^T \in \Re^{1 \times m}. \tag{3.3}$$

Hence

$$\mathbf{x}(n+j) = P_n^{(j+1)} \cdot \mathbf{x}(n-1) \quad \text{and} \quad y(n+j) = C \cdot P_n^{(j+1)} \cdot \mathbf{x}(n-1), \tag{3.4}$$

where

$$P_n^{(j+1)} \doteq \prod_{i=0}^{j} A(n+i) \doteq A(n+j)A(n+j-1)\cdots A(n+1)A(n). \tag{3.5}$$

Now, for a given $j = 0, 1, 2, \ldots$, if

$$\|P_n^{(j+1)}\| \leq \gamma < 1, \ \forall n, \tag{3.6}$$

then, $lim_{n\to\infty} y(n) = 0$ implying AS of (3.2) (or (2.1)). Here, $\|\cdot\|$ denotes any mutually consistent matrix norm [14], in particular, the p-norms.

Remarks:

1) Note that

$$P_n^{(j+1)} = A(n+j) \cdot P_n^{(j)}, \ \forall n. \tag{3.7}$$

2) Note that $A(n+j)$ is the corresponding system matrix in its companion form as in (3.3). Taking this structure into account, one notices that premultiplication of $P_n^{(j)}$ by $A(n+j)$ simply shifts the last $n-1$ rows of $P_n^{(j)}$ upwards by one row. Hence, the first $n-1$ rows of $P_n^{(j+1)}$ are identical to the last $n-1$ rows of $P_n^{(j)}$.

3) With remark 2 in mind, it is not difficult to show that

$$P_n^{(j)} = \begin{bmatrix} 0_{(m-j)\times j} & & \mathbf{I}_{(m-j)\times(m-j)} \\ & \left\{p_{\alpha,\beta}^{(j)}\right\}_{\substack{\alpha=m-j+1,\ldots,m \\ \beta=1,\ldots,m}} & \end{bmatrix},$$

$$j = 1, \ldots, m, \tag{3.8}$$

where $P_n^{(j)} \doteq \left\{p_{\alpha,\beta}^{(j)}\right\}_{\alpha,\beta=1,\ldots,m} \in \Re^{m\times m}$.

Using the ∞-norm, the condition for AS in (3.6) may now be utilized to prove Theorem II.1. First, we need

Lemma III.1: Consider the TV system in (3.2)-(3.3). Whenever $\mathbf{a}(n) \in \Omega$, $\forall n$,

$$\begin{aligned} \|P_n^{(m)}\|_\infty &= \| \prod_{i=0}^{m-1} A(n+i)\|_\infty \\ &= \|A(n+m-1)A(n+m-2)\cdots A(n+1)A(n)\|_\infty \\ &\leq \gamma < 1, \ \forall n. \end{aligned}$$

Here, Ω is the region given by Theorem II.1.

Proof: What we need to ascertain is that, whenever $\mathbf{a}(n) \in \Omega$, $\forall n$, the product of m consecutive $A(n+i)$'s, that is, $P_n^{(m)}$, $\forall n$, has an ∞-norm of not more than γ. With remark 2 above in mind, we show that, given $\mathbf{a}(n) \in \Omega$, $\forall n$, the newly computed last row of $P_n^{(j+1)}$ has an ∞-norm of not more than γ. If this holds true for m consecutive products, we will have $\|P_n^{(m)}\|_\infty \leq \gamma < 1$, as desired.

We proceed with an inductive scheme on $j = 1, 2, \ldots, m-1$. First, when $j = 1$, clearly the claim is true.

Next, assume that, for some $\ell = 1, 2, \ldots, m-1$, the last row of $P_n^{(\ell)}$ has an ∞-norm of not more than γ. Noting that $P_n^{(\ell)}$ is arrived at by ℓ consecutive multiplications, each row of the submatrix $\left\{p_{\alpha,\beta}^{(\ell)}\right\}_{\substack{\alpha=m-\ell+1,\ldots,m \\ \beta=1,\ldots,m}}$ must have an ∞-norm of not more than γ.

Finally, we need to show that, the newly computed last row of $P_n^{(\ell+1)}$ also has an ∞-norm of not more than γ. Note that, from (3.8), elements of this last row are given by

$$p_{m,\beta}^{(\ell+1)} = \begin{cases} \sum_{i=1}^{\ell} a_i p_{m+1-i,\beta}^{(\ell)}, & \text{for } \beta = 1, \ldots, \ell; \\ \sum_{i=1}^{\ell} a_i p_{m+1-i,\beta}^{(\ell)} + a_{m+1+\ell-\beta}, & \text{for } \beta = \ell+1, \ldots, m. \end{cases}$$

Hence, the ∞-norm of this last row is given by the equations shown at the bottom of the next page. This completes the proof.

Corollary III.2: Whenever $\mathbf{a}(n) \in \Omega, \forall n$, the TV system in (3.2) is AS.

Proof: This is immediate from Lemma III.1 and (3.6).

Remark: Note that,

$$\|P_n^{(i)}\|_\infty \geq 1 \quad \text{for} \quad i \leq m-1. \tag{3.9}$$

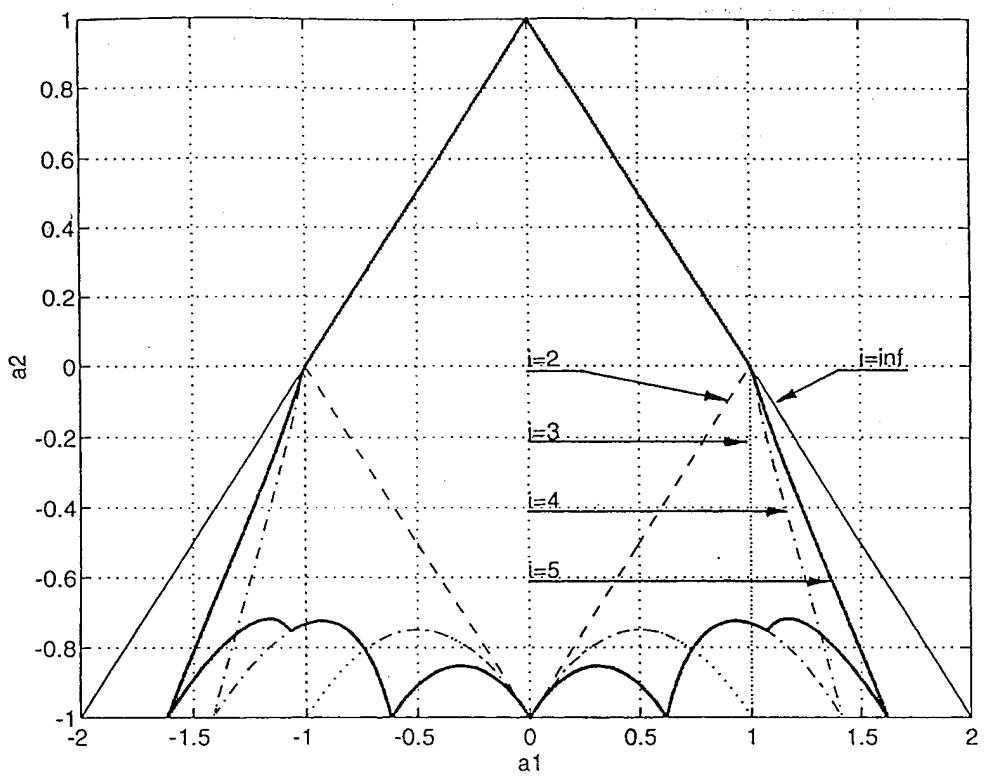


Fig. 1. The regions $\hat{\Omega}_{\infty,1}^{(i)}$, $i = 2,3,4,5,\infty$, for a second-order system.

Hence, if the norm condition in (3.6) is to be utilized for AS investigations, it is necessary to deal with at least m consecutive multiplications. The result in [10] utilizes *exactly* m such multiplications (see Lemma III.1).

Now, consider the following TI counterpart of the system in (3.2)–(3.3):

$$\hat{\mathbf{x}}(n) = \hat{A} \cdot \hat{\mathbf{x}}(n-1); \qquad \hat{y}(n) = \hat{C} \cdot \hat{\mathbf{x}}(n), \tag{3.10}$$

where

$$\hat{A} = \begin{bmatrix} 0 & 1 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \\ \hat{a}_m & \hat{a}_{m-1} & \cdots & \hat{a}_2 & \hat{a}_1 \end{bmatrix} \in \Re^{m\times m};$$

$$\hat{C} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}^T \in \Re^{1\times m}. \tag{3.11}$$

This TI system is AS if

$$|\lambda_i[\hat{A}]| < 1, \ \forall i = 1,\ldots,m. \tag{3.12}$$

Hence, the corresponding AS region is $\hat{\Omega}_1^{(\infty)}$ where

$$\hat{\Omega}_\omega^{(\infty)} = \left\{\hat{\mathbf{a}} \in \Re^m : |\lambda_i[\hat{A}]| < \omega, \ \forall i = 1,\ldots,m\right\}, \tag{3.13}$$

where $\omega \in \Re$. On the other hand, for $i = \aleph_+$, let

$$\hat{\Omega}_{p,\omega}^{(i)} = \left\{\hat{\mathbf{a}} \in \Re^m : \|\hat{P}^{(i)}\| < \omega\right\}, \tag{3.14}$$

where $\hat{P}^{(i)} \doteq \hat{A}^i$. Note that, in general,

$$\hat{\Omega}_{p,1}^{(i)} \subset \hat{\Omega}_1^{(\infty)}. \tag{3.15}$$

However, since $\lim_{i\to\infty} \hat{A}^i = 0_{m\times m}$ for $\hat{\mathbf{a}} \in \hat{\Omega}_{p,1}^{(\infty)}$, we have

$$\lim_{i\to\infty} \hat{\Omega}_{p,1}^{(i)} = \hat{\Omega}_1^{(\infty)}. \tag{3.16}$$

In fact, although $\hat{\Omega}_{p,\omega}^{(i)} \not\supset \hat{\Omega}_{p,\omega}^{(i-1)}$, typically, $\hat{\Omega}_{p,\omega}^{(i)}$ is larger in "volume" than $\hat{\Omega}_{p,\omega}^{(i-1)}$. For $m = 2$, Fig. 1 shows $\hat{\Omega}_{\infty,1}^{(i)}$, for $i = 1,2,3,4,5$, where $\hat{\Omega}_1^{(\infty)}$, the familiar triangular stability region, is also shown. Similar figures for other norms may also be obtained.

$$\begin{aligned}\sum_{\beta=1}^{m}\left|p_{m,\beta}^{(\ell+1)}\right| &= \sum_{\beta=1}^{\ell}\left|\sum_{i=1}^{\ell} a_i p_{m+1-i,\beta}^{(\ell)}\right| + \sum_{\beta=\ell+1}^{m}\left|\sum_{i=1}^{\ell} a_i p_{m+1-i,\beta}^{(\ell)} + a_{m+1+\ell-\beta}\right| \\ &\le \sum_{i=1}^{\ell}|a_i|\sum_{\beta=1}^{\ell}\left|p_{m+1-i,\beta}^{(\ell)}\right| + \sum_{i=1}^{\ell}|a_i|\sum_{\beta=\ell+1}^{m}\left|p_{m+1-i,\beta}^{(\ell)}\right| + \sum_{\beta=\ell+1}^{m}|a_{m+1+\ell-\beta}| \\ &< \sum_{i=1}^{\ell}|a_i| + \sum_{i=\ell+1}^{m}|a_i| = \sum_{i=1}^{m}|a_i| \le \gamma < 1.\end{aligned}$$

Now, it is instructive to see whether a norm condition on $P_n^{(i)}$, $i > m$, may provide a larger region of AS. Given $A(n)\Re^{m\times m}$, let us denote the perturbation on $a_i(n) \in \Re$ at time instant n by $\Delta_i(n) \in \Re$. Taking

$$\Delta(n) = \begin{bmatrix} & \mathbf{0}_{(m-1)\times m} & \\ \Delta_m(n) & \Delta_{m-1}(n) \quad \cdots \quad \Delta_2(n) & \Delta_1(n) \end{bmatrix} \in \Re^{m\times m}, \tag{3.17}$$

we now have

$$\begin{aligned} A(n+1) &= A(n) + \Delta(n); \\ A(n+2) &= A(n+1) + \Delta(n+1) = A(n) + \Delta(n+1) + \Delta(n) \\ &\vdots \\ A(n+j) &= A(n) + \sum_{k=0}^{j-1} \Delta(n+k). \end{aligned} \tag{3.18}$$

Hence

$$P_n^{(j+1)} = \prod_{i=0}^{j} A(n+i) = \prod_{i=0}^{j} \left(A(n) + \sum_{k=0}^{i-1} \Delta(n+k) \right) \tag{3.19}$$

$$= \hat{P}_n^{(j+1)} + \Delta_n^{(j+1)}, \tag{3.20}$$

where

$$\hat{P}_n^{(j+1)} \doteq A(n)^{j+1} \in \Re^{m\times m} \tag{3.21}$$

is obtained by consecutive multiplication of the *same* system matrix $A(n)$ (compare with $P_n^{(j+1)}$) and $\Delta_n^{(j+1)} \in \Re^{m\times m}$ is a certain matrix that is solely due to the coefficient perturbations. In fact,

$$\lim_{\substack{\Delta_i \to 0 \\ \forall i=1,\ldots,m}} \Delta_n^{(j+1)} = \mathbf{0}_{m\times m}. \tag{3.22}$$

It is (3.20) that we will utilize to arrive at our goal.

To satisfy the norm condition in (3.6), we need

$$\|\hat{P}_n^{(j+1)} + \Delta_n^{(j+1)}\| \le \gamma < 1. \tag{3.23}$$

To ensure (3.23), let $\|\hat{P}_n^{(j+1)}\| + \|\Delta_n^{(j+1)}\| \le \gamma < 1$, which is satisfied if the following two conditions are valid:

(c1) Choose $\mathbf{a}(n) \in \Re^m$ such that

$$\|\hat{P}_n^{(j+1)}\| < \delta < \gamma, \quad \text{that is,} \quad \mathbf{a}(n) \in \hat{\Omega}_{p,\delta}^{(j+1)}, \ \forall n. \tag{3.24}$$

(c2) Choose the perturbation allowed on each coefficient such that

$$\|\Delta_n^{(j+1)}\| < 1 - \delta. \tag{3.25}$$

It is not hard to see that, in general, $\Delta_n^{(j+1)}$ is a function of $\Delta(n+i)$, $\forall i = 0,\ldots,j-1$, and $A(n)^i$, $i = 1,\ldots,j$. Hence, it may be expressed as

$$\Delta_n^{(j+1)} = \mathbf{f}(\Delta(n),\ldots,\Delta(n+j-1); A(n),\ldots,A(n)^j). \tag{3.26}$$

Taking the maximum allowable rate of change at each time instant to be equal, let

$$\|\Delta(n+i)\| \le \|\Delta\|, \ \forall i = \aleph_+^0. \tag{3.27}$$

Then, we may obtain a bound for $\|\Delta_n^{(j+1)}\|$ as a function of $\|\Delta\|^i$ and $\|A(n)^i\|$, $\forall i = 1,\ldots,j$, that is,

$$\|\Delta_n^{(j+1)}\| \le g(\|\Delta\|,\ldots,\|\Delta\|^j; \|A(n)\|,\ldots,\|A(n)^j\|). \tag{3.28}$$

Next, in order to enforce a bound on g, within $\hat{\Omega}_{p,\delta}^{(j+1)}$, find the maximum values of $\|A(n)^i\|, i = 1,\ldots,j$. For this, one needs to perform a search over the $(m-1)$-dimensional boundary of $\hat{\Omega}_{p,\delta}^{(j+1)}$. When these maximum values are substituted, the bound for g obtained thus is a polynomial of degree j in $\|\Delta\|$ with no constant coefficient. Hence, let this bound be

$$h \doteq \sum_{i=0}^{j} h_i \|\Delta\|^i < 0, \ h_0 \doteq -(1-\delta), \ 0 < \delta < 1. \tag{3.29}$$

Clearly, $\|\Delta\| = 0$ (that is, the TI case) satisfies (3.29). Let r_i, $i = 1,\ldots,j$, be the non-negative (we are only interested in this case since $\|\Delta\| \ge 0$) roots of $h = 0$. Note that, 0 cannot be a root of h. If $0 < r_1 < r_2 < \cdots < r_j$, we conclude that $0 \le \|\Delta\| < r_1$. Note that, $r_2 < \|\Delta\| < r_3$ is of no use since, at any arbitrary instant in time, we must allow for the coefficients to be stationary, that is, $\|\Delta\| = 0$.

Summarizing the above, assuming (3.27), we have the following:

Step I: Pick $j = 2$. With $j = 1$, Theorem II.1 provides the "optimal" region. For a suitable $0 < \delta < 1$, let $\mathbf{a}(n) \in \hat{\Omega}_{p,\delta}^{(j+1)}$, $\forall n$.

Step II: Find the bounding function g in (3.28). An appropriate algorithm that is applicable to a system of any general order is in [15].

Step III: On the $(m-1)$-dimensional boundary of $\hat{\Omega}_{p,\delta}^{(j+1)}$, find the maximum values of $\|A(n)^i\|, i = 1,\ldots,j$.

Step IV: Obtain h in (3.29).

Step V: The maximum allowable rate of change of coefficients within $\hat{\Omega}_{p,\delta}^{(j+1)}$ is the least positive root of h.

Step VI: If the actual allowable rate of change is higher, one may repeat the procedure with a lower δ resulting in a smaller region. If the actual allowable rate of change is restricted to be lower, one may repeat the procedure with a higher j resulting in a larger region.

Remarks:

1) Step III can be very computer intensive. For second-order systems, however, this search procedure is quite easy. See next section.
2) By using different norms, one may obtain different "shapes" of regions (in the coefficient space) for the maximum allowable perturbation on the coefficients. For example, ∞-norm gives a diamond-shaped region; 2-norm gives a circular region; 1-norm gives a box-shaped region.

V. Example

Our main results are now illustrated through an example. Due to the light computer burden and the possibility of graphically representing the relevant regions, we consider a TV system of order 2.

Step I: Let $j = 2$. The regions $\hat{\Omega}_{\infty,\delta}^{(3)}$ for $\delta = 0.5, 0.6, 0.7, 0.8, 0.9, 1.0$ are in Fig. 2.

Step II: Note that, $P_n^{(3)} = (A(n) + \Delta(n) + \Delta(n+1))(A(n) + \Delta(n))A(n) = \hat{P}_n^{(3)} + \Delta_n^{(3)}$, where $\hat{P}_n^{(3)} = A(n)^3$ and

$$\begin{aligned} \Delta_n^{(3)} &= (\Delta(n) + \Delta(n+1))(A(n) + \Delta(n))A(n) \\ &\quad + A(n)\Delta(n)A(n) \\ &= \Delta(n)A(n)^2 + \Delta(n)^2 A(n) + \Delta(n+1)A(n)^2 \\ &\quad + \Delta(n+1)\Delta(n)A(n) + A(n)\Delta(n)A(n). \end{aligned}$$

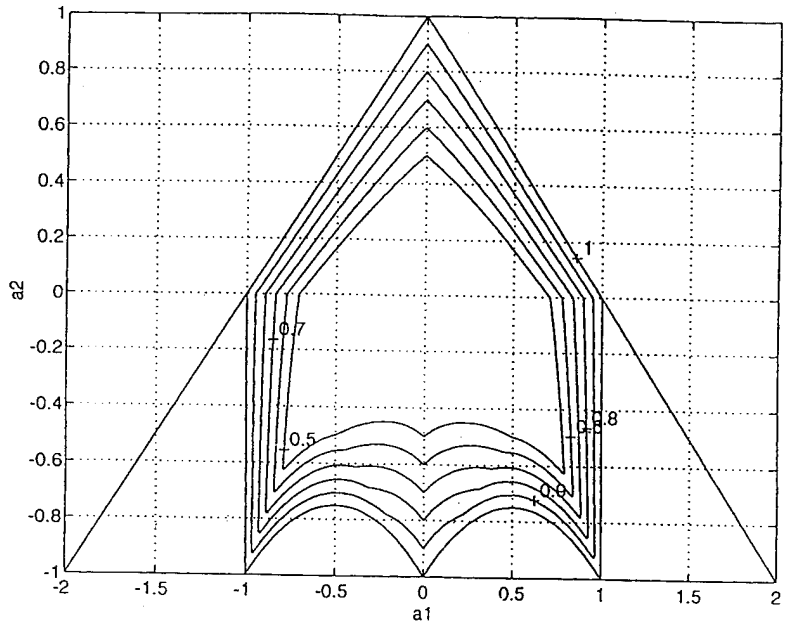


Fig. 2. The regions $\hat{\Omega}^{(3)}_{\infty,\delta}$, $\delta = 0.5, 0.6, 0.7, 0.8, 0.9, 1$, for a second-order system.

TABLE I
$\|\Delta\|_{\infty,max}$ OBTAINED WITH $\hat{\Omega}^{(3)}_{\infty,\delta}$ REGIONS FOR A SECOND-ORDER SYSTEM

δ	$\|A(n)\|_{\infty,\max} = \|A(n)^2\|_{\infty,\max}$	$\|\Delta\|_{\infty,\max}$
0.900	1.886	0.014
0.800	1.777	0.029
0.700	1.664	0.048
0.600	1.547	0.070
0.500	1.415	0.098

TABLE II
$\|\Delta\|_{\infty,max}$ OBTAINED WITH $\hat{\Omega}^{(4)}_{\infty,\delta}$ REGIONS FOR A SECOND-ORDER SYSTEM

δ	$\|A(n)\|_{\infty,\max} = \|A(n)^2\|_{\infty,\max}$	$\|A(n)^3\|_{\infty,\max}$	$\|\Delta\|_{\infty,\max}$
0.900	2.312	2.232	0.004
0.800	2.213	2.063	0.009
0.700	2.121	1.906	0.015
0.600	2.011	1.725	0.022
0.500	1.873	1.517	0.031

The last equality above is **f** in (3.26). Assuming (3.27), g in (3.28) is

$$\begin{aligned}\|\Delta_n^{(3)}\|_\infty &\le g(\|\Delta\|_\infty, \|\Delta\|_\infty^2, \|A(n)\|_\infty, \|A(n)^2\|_\infty)\\ &= 2\cdot\|\Delta\|_\infty\cdot\|A(n)^2\|_\infty + 2\cdot\|\Delta\|_\infty^2\cdot\|A(n)\|_\infty\\ &\quad + \|\Delta\|_\infty\cdot\|A\|_\infty^2.\end{aligned}$$

Step III: For the second-order case consider here, it is extremely easy to find the corresponding maximum values of $\|A(n)\|_\infty$ and $\|A(n)^2\|_\infty$. In fact, these simultaneously occur at the same point. See Table I.

Step IV: The function h in (3.29) is given by

$$\begin{aligned}h = [2\cdot\|A\|_{\infty,\max}]\cdot\|\Delta\|_\infty^2 + [2\|A^2\|_{\infty,\max} + \|A\|_{\infty,\max}^2]\\ \cdot\|\Delta\|_\infty - (1-\delta).\end{aligned}$$

Step V: The lowest positive root of h gives the maximum allowable rate of change of coefficients within $\hat{\Omega}^{(3)}_{\infty,\delta}$. See Table I.

Step VI: The procedure above was repeated for $j = 3$. The regions $\hat{\Omega}^{(4)}_{\infty,\delta}$ are in Fig. 3 while the corresponding results are

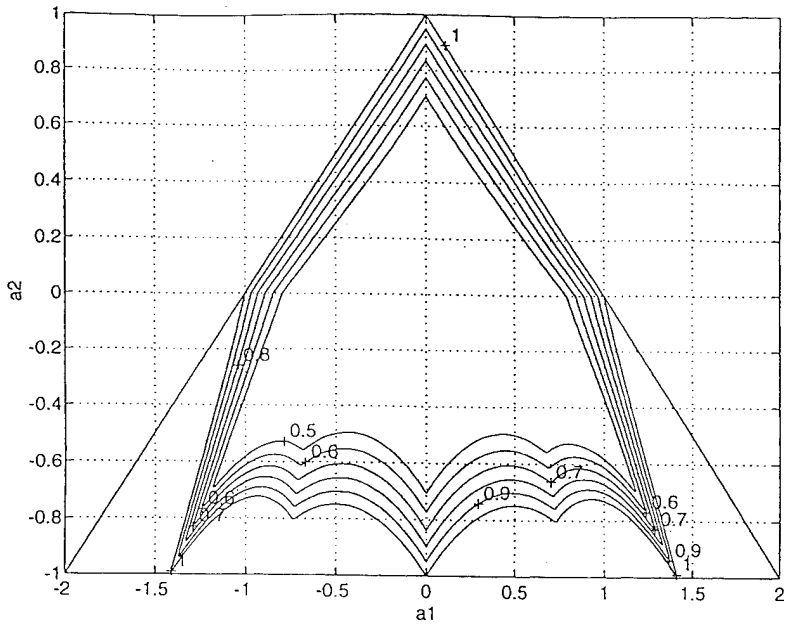


Fig. 3. The regions $\hat{\Omega}^{(4)}_{\infty,\delta}$, $\delta = 0.5, 0.6, 0.7, 0.8, 0.9, 1$, for a second-order system.

tabulated in Table II. Note how it verifies the remarks made under Step VI in Section III.

Remarks:

1) Similar computations may be done using the 2- and 1-norms as well.
2) The value of $\|\Delta\|_{\infty,\max}$ obtained with the use of the ∞-norm provides a diamond-shaped region at *each* coefficient $\mathbf{a}(n) \in \hat{\Omega}^{(j+1)}_{\infty,\delta}$ where it may be perturbed to $\mathbf{a}(n+1) \in \hat{\Omega}^{(j+1)}_{\infty,\delta}$. Similarly, $\|\Delta\|_{2,\max}$ corresponds to a circular region while $\|\Delta\|_{1,\max}$ corresponds to a box-shaped region.
3) Often, the *coefficients* are given to be TV and restricted to be within a box-shaped region in the coefficient space. Note the difference between the coefficients and the perturbations. Item 2 above describes the "shape" of possible *perturbations*. Such a situation may be easily incorporated into the above procedure to obtain a value for the maximum rate of change that is sufficient for AS.

VI. Conclusion and Final Remarks

The results presented in this paper provides a technique that may be utilized to obtain regions of AS in the coefficient space of linear TV difference equations. The regions thus obtained incorporate information regarding the maximum rate of change of the system parameters.

The technique proposed yield only sufficient conditions. In dealing with higher order systems, the computational burden may be quite heavy. However, the regions $\hat{\Omega}^{(j+1)}_{p,\delta}$ are invariant for a given m, and hence, it is only necessary to compute them once. When $m = 2$, the application of the proposed method is quite straightforward. It is the authors' hope that this work will encourage improvements to the technique presented and development of alternate algorithms.

Acknowledgment

The fruitful discussions with Professors Eliahu I. Jury, of University of Miami, and Peter H. Bauer, of University of Notre Dame, are gratefully appreciated. The second author wishes to express gratitude for the support received from the Department of Electrical and Computer Engineering, University of Miami, where he was a visiting scholar while this work was being performed.

References

[1] V. L. Kharitonov, "Asymptotic stability of an equilibrium position of a family of systems of linear differential equations," *Differential'nye Uravneniya,* vol. 14, pp. 1483–1485, 1979.
[2] M. Mansour, S. Balemi, and W. Truöl, Eds., *Robustness of Dynamic Systems with Parameter Uncertainties.* Berlin: Birkhäuser, 1992.
[3] E. I. Jury, "Robustness of a discrete system," *Automatikha i Telemechanika,* vol. 51, no. 5, part 1, pp. 3–28, May 1990 (in Russian); also in *Automat. Remote Control,* pp. 571–592, Oct. 10, 1990 (in English).
[4] C. A. Desoer, "Slowly varying discrete systems $\mathbf{x}_{i+1} = \mathbf{A}_i\mathbf{x}_i$," *Electron. Lett.,* vol. 6, pp. 339–340, 1970.
[5] T. Mori, N. Fukuma, and M. Kuwahara, "A stability criterion for linear time-varying systems," *Int. J. Control,* vol. 34, pp. 585–591, 1981.
[6] E. W. Kamen, "The poles and zeros of a linear time-varying system," *Linear Algebra and Its Applications,* vol. 98, pp. 263–289, 1988.
[7] J.-H. Su and I.-K. Fong, "New robust stability bounds of linear discrete-time systems with time-varying uncertainties," *Int. J. Control,* vol. 58, pp. 1461–1467, 1993.
[8] F. Amato, G. Celentano, and F. Garofalo, "New sufficient conditions for the stability of slowly varying linear systems," *IEEE Trans. Automat. Control,* vol. 38, pp. 1409–1411, Sept. 1993.
[9] P. H. Bauer and K. Premaratne, "Robust stability of time-variant interval matrices," in *Proc. 29th IEEE Conf. Decision and Control,* Honolulu, HI, Dec. 1990, pp. 334–335.
[10] P. H. Bauer, M. Mansour, and J. Durán, "Stability of polynomials with time-variant coefficients," *IEEE Trans. Circuits Syst.—I,* vol. 40, pp. 423–426, June 1993.
[11] P. H. Bauer, K. Premaratne, and J. Durán, "A necessary and sufficient condition for robust stability of time-variant discrete systems," *IEEE Trans. Automat. Control,* vol. 38, pp. 1427–1430, Sept. 1993.
[12] P. H. Bauer, personal communication, 1994.
[13] J. Jiang, "Design of reconfigurable control systems using eigenstructure assignments," *Int. J. Control,* vol. 59, pp. 395–410, 1994.
[14] G.H. Golub and C.F. Van Loan, *Matrix Computations.* Baltimore, MD: Johns Hopkins, 1983.
[15] K. Premaratne and M. Mansour, "Robust stability of time-variant discrete-time systems with bounded parameter perturbations," Internal Report TR-94-03, Dept. of Electrical and Computer Engineering, Univ. of Miami, Coral Gables, FL, 1994.

IEEE Transactions on Signal Processing, in review, 1995

Two-Channel IIR QMF Banks with Approximately Linear-Phase Analysis and Synthesis Filters

Mahes M. Ekanayake,[2] *Student Member IEEE* **and Kamal Premaratne,** *Senior Member IEEE*

Abstract- **Perfect linear-phase two-channel QMF banks require the use of FIR analysis and synthesis filters. Although they are less expensive and yield superior stopband characteristics, perfect linear-phase cannot be achieved with stable IIR filters. Thus, IIR designs usually incorporate a postprocessing equalizer which is optimized to reduce the phase distortion of the entire filter bank. However the analysis and synthesis filters of such a IIR filter bank are not linear-phase. In this paper, a computationally simple method to obtain IIR analysis and synthesis filters that possess negligible phase distortion is presented. The method is based on first applying the balanced reduction procedure to obtain nearly all-pass IIR polyphase components and then approximating these with perfect all-pass IIR polyphase components. The resulting IIR designs already have only negligible phase distortion. However, if required, further improvement may be achieved through optimization of the filter parameters. For this purpose, a suitable objective function is presented. Bounds for the magnitude and phase errors of the designs are also derived. Design examples indicate that the derived IIR filter banks are more efficient in terms of computational complexity than the FIR prototypes. Simulations show that the IIR filters perform better than FIR Perfect Reconstruction systems under coefficient quantization.**

1 Introduction

A maximally decimated [1] two-channel Quadrature Mirror Filter (QMF) bank is shown in Fig. 1, where $H_0(z)$, $H_1(z)$ are the transfer functions of the analysis filters and $F_0(z)$,$F_1(z)$ are the synthesis filters [2]. The analysis bank separates the signal into two half-band signals and the synthesis bank reconstructs the signal from the two half-band signals. To achieve good separation between the two half-band signals the stop-band energies of the two analysis filters have to be minimized. The amount of stopband energy that can be tolerated will depend upon the application. For instance in subband coding of speech [4], the spectral energy across a frequency range of $300 - 3000\ Hz$ may exhibit a difference of 40 dB. Hence the analysis filters must have a stopband attenuation of atleast 40 dB. However with a frequency range of $100 - 6900\ Hz$, the required stopband attenuation may be as much as $60 - 70\ dB$.

The reconstructed signal in general suffers from aliasing distortion (ALD), amplitude distortion (AMD)

[2]Corresponding author

The authors greatfully acknowledge the support received from the ONR Grant N00014-94-1-0454. The authors are with the Department of Electrical and Computer Engineering, P.O. Box 248294, University of Miami, Coral Gables, FL 33124. Phone: 305-284-3291 Fax: 305-284-4044 E-mail: mahes@obsidian.eng.miami.edu

and phase distortion (PHD) due to the fact that the analysis and synthesis filters are not ideal. Hence a common requirement in most applications is that the reconstructed signal, $\hat{x}(n)$ be as close to $x(n)$ as possible. However other constraints are usually imposed to reduce nonlinear distortions, such as coding errors and transmission channel distortions, that cannot be directly evaluated. One such constraint that is usually imposed is that the analysis/synthesis filters be linear-phase [2]. In particular, such a constraint is typically imposed in digital audio applications.

Several techniques for the design of linear phase finite impulse response (FIR) filters which eliminate ALD and PHD while minimizing AMD have been reported [4][5]. We shall call these *Type I systems.* We shall show that for a Type I system, to eliminate AMD completely, it is required that the polyphase components must be all-pass (AP). However since it is not possible to have FIR AP filters (excepting the trivial case of a delay), it is only possible to minimize AMD by optimization. Techniques for the design of FIR filters which have the perfect reconstruction (PR) property, where all ALD, AMD and PHD are eliminated have also been reported [6][7][8]. It is also possible to incorporate the linear-phase property into these filters [9]. We shall call such a PR system with the linear-phase property a *Type II system.* However the stop-band energy of a Type II filter is much greater than that of a Type I filter of the same length. It turns out that to achieve the same stop-band energy with a Type II filter, it would require approximately twice the length of a Type I filter. Hence the group delay of a Type II system is approximately twice that of a comparable Type I system. The number of multiplications per unit time (MPU's) for a Type II system (of twice the length of a Type I system) is the same as that for a Type I system, but the number of additions per unit time (APU's) are much more [9].

Although structurally robust Type II systems can be implemented using the two-multiplier lattice structures proposed in [9], the computational complexity of these are twice that of the most efficient one-multiplier lattice implementation proposed in [3]. The two-multiplier lattice structures are robust due to the fact that the analysis and synthesis banks have coefficients with the same magnitude but opposite sign. However the one-multiplier lattice structures are not structurally robust PR systems since the coefficients in the analysis and synthesis banks are different. It turns out that the Type II systems, when implemented with the one-multiplier lattice structure, are highly sensitive to coefficient quantization as illustrated by the examples in section 5.

It is well known that an infinite impulse response (IIR) filter which has the same stop-band energy as that of a FIR filter will be of much lower order and hence is less expensive and more efficient computationally. However IIR filters are seldom used for filter bank systems due to the fact that they inherently produce phase distortion. Several IIR filter banks which have no ALD and AMD have been reported [11][12][13][14]. The PHD is minimized by using a separate AP equalizer network once the signal is reconstructed [2]. Hence these IIR filter banks are not suitable for applications such as digital audio, where the linear-phase property of the analysis and synthesis filters are desired. In [10] a method based on the eigenfilter approach has been proposed for the design of IIR filter banks with an approximately linear phase response in the pass band of the analysis and synthesis banks. Although this method yields filters with very low stopband energy, the overall PHD is not minimized and hence

the filters exhibit a high amount of PHD. An alternative approach to design approximately linear phase IIR filter banks that has not been investigated so far is to eliminate ALD and AMD and minimize PHD.

In this paper, a computationally and numerically efficient method for the design of IIR filter banks, where AMD and ALD are eliminated and PHD is minimized is presented. Each of the analysis and synthesis filters are designed to have negligible phase distortion. The method which is based on approximating a FIR filter with an IIR filter consists of the following steps.

1. Design a *suitable* linear phase FIR prototype filter. Such a filter must have polyphase components which are approximately AP.
2. Obtain an IIR filter bank having the same magnitude and phase responses (approximately) as that of the FIR prototypes by the application of the balanced reduction (BR) procedure.
3. Optimize the parameters of the IIR filter bank to get the best IIR approximation. Here, the main objective is to reduce the phase distortion as far as possible.

At this point, we must mention that it is possible to do a direct optimization without using the BR procedure. However this method turns out to be computationally inefficient compared to the proposed method. A comparison of the time taken for the two methods is given in section 5.

The organization of the paper is as follows. In section 2, a brief overview of the design procedure of a suitable FIR prototype filter bank is presented. In section 3, a brief review of the BR algorithm together with some new results applicable to the task at hand are discussed. In section 4, a suitable objective function and its application for the optimization procedure are presented. Finally, in section 5, application examples, effects of coefficient quantization and roundoff noise, and a comparison with FIR filter banks are given.

2 Design of Linear-Phase FIR Filter Banks

In IIR filter banks, it is desirable to have the polyphase components as AP filters, since with this choice, AMD is eliminated [1] (see also equation (5)). Moreover, AP filters can be implemented very efficiently. With this in mind, the most suitable FIR prototype filter bank is the Type I FIR filter bank. As we shall show, for this class of filters, the polyphase components turn out to be nearly AP. This observation is crucial in approximating the prototype filter with true AP filters with no appreciable error. In general the polyphase components of Type II systems are not approximately AP.

The design of Type I FIR filter banks involves the minimization of AMD by optimizing the filter coefficients. Such optimization has been done by Johnston [4], and Jain and Crochiere [5]. We discuss briefly the method due to Johnston and show that the polyphase components are approximately AP.

For the filter bank in Fig. 1, the reconstructed signal is given by

$$\hat{X}(z) = \frac{1}{2}[H_0(z)F_0(z) + H_1(z)F_1(z)]X(z) + \frac{1}{2}[H_0(-z)F_0(z) + H_1(-z)F_1(z)]X(-z) \qquad (1)$$

The second term represents the ALD which must be eliminated. Also in this method we design only the filter $H_0(z)$ and we choose

$$H_1(z) = H_0(-z) \tag{2}$$

$$F_0(z) = H_0(z) \qquad \text{and} \qquad F_1(z) = -H_0(-z) \tag{3}$$

The choice of (2) ensures that $H_1(z)$ is high-pass if $H_0(z)$ is low-pass. The choice of (3) ensures that ALD is eliminated and also efficient implementation of the filter bank is facilitated. With this choice, if we force $H_0(z)$ to be linear-phase then all the filters will be linear-phase and thus eliminate PHD. We can represent $H_0(z)$ in terms of its polyphase components [2] as

$$H_0(z) = E_0(z^2) + z^{-1}E_1(z^2) \tag{4}$$

The resulting polyphase representation of the filter bank is shown in Fig.2. Now with ALD eliminated, the transfer function of the entire filter bank becomes

$$T(z) = \frac{\hat{X}(z)}{X(z)} = 2z^{-1}E_0(z^2)E_1(z^2) \tag{5}$$

Since $H_0(z)$ is linear phase it has to be of even length [1]. Hence it can be easily verified that

$$|E_0(e^{j\omega})| = |E_1(e^{j\omega})| \tag{6}$$

Hence, to completely eliminate AMD, we must have,

$$|E_0(e^{j2\omega})| = |E_1(e^{j2\omega})| = k \tag{7}$$

where k is a real constant. In particular, we choose $k = \frac{1}{2}$ so that $H_0(1) = 1$. Keeping $E_0(z)$ and $E_1(z)$ to be FIR, due to the lack of AP property, this condition cannot be satisfied. In a typical design of $H_0(z)$ we first force it to be linear-phase and then optimize the parameters such that (7) holds approximately and the stop band energy of $H_0(z)$ is minimum. The objective function that has been used for this is [4]

$$\alpha \int_{\omega_s}^{\pi} |H_0(e^{j\omega})|^2 d\omega + \int_0^{\pi/2} [|H_0(e^{j\omega})|^2 + |H_1(e^{j\omega})|^2 - 1]^2 d\omega \tag{8}$$

where ω_s is the stop band edge. It can be shown that due to (6) , the objective function can be simplified to

$$\alpha \int_{\omega_s}^{\pi} |H_0(e^{j\omega})|^2 d\omega + \int_0^{\pi/2} [4|E_0(e^{j2\omega})|^2 - 1]^2 d\omega \tag{9}$$

Hence the filter parameters of the FIR filter are optimized such that (7) holds approximately. Therefore for the optimized filter, although (7) does not hold, the following is true.

$$|E_0(e^{j2\omega})| = |E_1(e^{j2\omega})| \approx \frac{1}{2} \tag{10}$$

This implies that the polyphase components of the FIR filter bank, $E_0(z)$ and $E_1(z)$ are only approximately AP. Note that due to (2) and (3), in the design of the FIR prototype filter bank we need only

to design $H_0(z)$.

In what follows, we shall approximate the FIR prototype filter $H_0(z)$ obtained as above with an IIR filter, $\tilde{H}_0(z)$. The polyphase components of $\tilde{H}_0(z)$ will be AP. The IIR filter bank is designed as a QMF bank. Hence the rest of the filters in the filter bank are obtained by relationships similar to (2) and (3). Hence AMD and ALD are eliminated. As we shall show, the approximation procedure ensures that the IIR filter $\tilde{H}_0(z)$ is approximately linear phase in the entire baseband ($0 \leq \omega \leq \pi$) provided that the FIR filter $H_0(z)$ has the linear phase property. Due to this we have shown [15] that PHD is also negligible.

3 The Balanced Reduction Procedure

In this section we discuss the Balanced Reduction (BR) procedure. As proposed by Moore [16], this is a very attractive procedure to derive a reduced order model from a given high order system. In this method, the given state space (s.s) formulation is transformed into a coordinate system wherein each state is as reachable as it is observable. This transformed system is called *balanced*, and by deleting the least reachable and observable states, a reduced model of the original results. For the purpose at hand, since the higher-order system is nearly AP, the lower order subsystem must also be nearly AP. We shall show that with the BR procedure, the nearly AP nature of the reduced order system enables us to find this dominant subsystem easily. Frequency error bounds for the approximation are also presented.

3.1 The Balanced Realization

The state-space realization of a system is not unique. The realization is minimal if it is both reachable and observable. Let (A, B, C, D) be a minimal s.s realization of a stable transfer function $H(z)$ of order m. The two positive definite matrices P and Q, which are called the reachability and observability gramians are defined as,

$$P = \frac{1}{2\pi j} \oint_{|z|=1} F(z)F^*(z)\frac{dz}{z} \tag{11}$$

$$Q = \frac{1}{2\pi j} \oint_{|z|=1} G^*(z)G(z)\frac{dz}{z} \tag{12}$$

$$\text{where} \quad F(z) = [zI - A]^{-1}B \tag{13}$$

$$\text{and} \quad G(z) = C[zI - A]^{-1} \tag{14}$$

P and Q can be found by solving the pair of Lyapunov equations given by

$$P - APA^T = BB^T \tag{15}$$

$$Q - A^TQA = C^TC \tag{16}$$

These can be solved efficiently using the Bartels-Stewart algorithm [17]. The Hankel singular values of the system are defined as the eigenvalues of the positive definite matrix PQ. A non-singular similarity transformation T of the state variables yields the similar system $(\hat{A}, \hat{B}, \hat{C}, D)$ where

$$\hat{A} = TAT^{-1}, \quad \hat{B} = TB, \quad \hat{C} = CT^{-1} \tag{17}$$

The transfer function and its Hankel singular values are invariant under a nonsingular similarity transformation. It is well known [16] that there exists a non-singular matrix T such that the similar system $(\hat{A}, \hat{B}, \hat{C}, D)$ obtained as in (17) is a balanced realization in the sense that the corresponding gramians are diagonal and identical, that is the reachability and observability gramians, $\hat{P}$ and $\hat{Q}$, of the system $(\hat{A}, \hat{B}, \hat{C}, D)$ takes the form

$$\hat{Q} = \hat{P} = \Sigma = diag(\sigma_1, \sigma_2, \ldots, \sigma_m) \tag{18}$$

Since the hankel singular values are invariant under a similarity transformation $\sigma_1^2 \geq \sigma_2^2 \geq \ldots \geq \sigma_m^2$ are the eigenvalues of PQ. Any realization satisfying (18) is called a balanced realization of $H(z)$. It must be noted that efficient and numerically stable methods to compute the balanced realization are available [18]. For example , see the routine 'dbalreal' in the MATLAB control system toolbox [20].

3.2 The Balanced Approximation of the Polyphase Components

The first step in the BR procedure is to obtain a balanced realization of $H(z)$. The key to the reduction procedure is the matrix Σ. Let Σ be decomposed into two parts;

$$\Sigma = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} \tag{19}$$

$$\Sigma_1 = diag(\sigma_1, \sigma_2, \ldots, \sigma_{\check{m}}) \tag{20}$$

$$\Sigma_2 = diag(\sigma_{\check{m}+1}, \sigma_{r+2}, \ldots, \sigma_m) \tag{21}$$

We can represent the balanced realization according to the partition (19) as

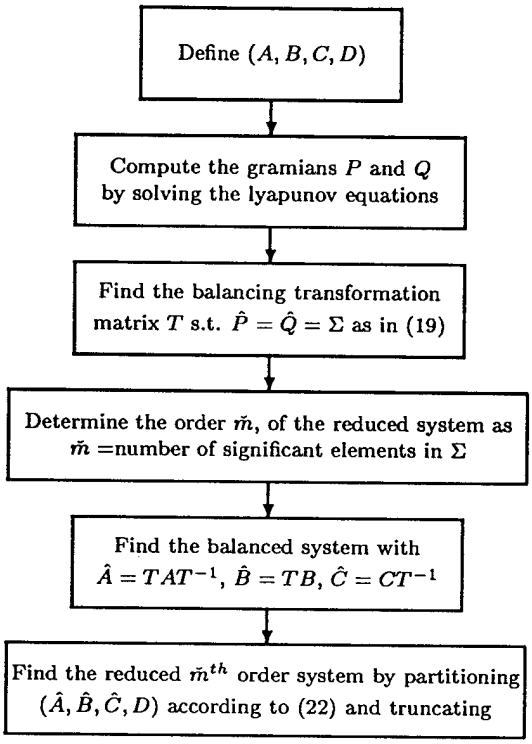
$$\hat{A} = \begin{bmatrix} \hat{A}_{11} & \hat{A}_{12} \\ \hat{A}_{21} & \hat{A}_{22} \end{bmatrix}, \quad \hat{B} = \begin{bmatrix} \hat{B}_1 \\ \hat{B}_2 \end{bmatrix}, \quad \hat{C} = \begin{bmatrix} \hat{C}_1 & \hat{C}_2 \end{bmatrix} \tag{22}$$

If $\sigma_{\check{m}} \gg \sigma_{\check{m}+1}$, then $(\hat{A}_{11}, \hat{B}_1, \hat{C}_1, D)$ represents the most observable and reachable part of $(\hat{A}, \hat{B}, \hat{C}, D)$. The system $(\hat{A}_{11}, \hat{B}_1, \hat{C}_1, D)$ is called the balanced approximation (BA) of the original system (A, B, C, D). The balanced approximation represents a good lower-order approximation of the original system if $\sigma_{\check{m}} \gg \sigma_{\check{m}+1}$. In fact, the following result is well known [21]. If $H(z)$ is the transfer function of the higher order system (A, B, C, D) and $\check{H}(z)$ is the transfer function of the BA $(\hat{A}_{11}, \hat{B}_1, \hat{C}_1, D)$, the frequency response error is bounded as

$$\|H(z) - \check{H}(z)\|_\infty \leq 2 \sum_{i=\check{m}+1}^{m} \sigma_i \tag{23}$$

where $\|.\|_\infty = sup_{\omega \epsilon \Re} |H(e^{j\omega})|$ and $\Re$ is the field of real numbers. Thus if $\sigma_{\check{m}} \gg \sigma_{\check{m}+1}$ we can expect a good approximation. It is also true that if $H(z)$ is stable then $\check{H}(z)$ is stable.

The application of the BR procedure to obtain a reduced order system from a higher order one is illustrated by the flowchart below.



Flowchart for balanced reduction procedure.

For the task at hand, we have a high order approximately AP FIR filter and our aim is to find a reduced order IIR model which is nearly AP. With this in mind we prove the following theorem:

Theorem 3.1 *Given a discrete-time state space realization (A, B, C) which is balanced, its Hankel singular values are all equal ($PQ = \sigma^2 I$) if and only if we can find a D such that the transfer function of (A, B, C, D), $H(z)$ is AP, i.e $|H(e^{j\omega})|^2 = \sigma^2$.*

Proof:

Let the discrete time system $H(z)$ have a s.s formulation (A, B, C). Let us use the bilinear transform $z = \frac{(1+s)}{(1-s)}$ to obtain the continuous time system $H_c(s)$ with s.s formulation (A_c, B_c, C_c) such that

$$H_c(s) = H(\tfrac{1+s}{1-s})$$

Then the proof follows from the following facts.

1. The theorem is true for continuous time systems [22].
2. The Hankel singular values of the discrete time system (A, B, C) are equal to that of the continuous time system (A_c, B_c, C_c) [22].

3. We can find a D such that (A, B, C, D) is AP if and only if there is a D_c such that (A_c, B_c, C_c, D_c) is AP. This follows from the fact that (A, B, C, D) is AP if and only if (A_c, B_c, C_c, D_c) is AP.

■

In [19] it has been proved that an orthogonal realization ($P = Q = I$) exists for an allpass filter. However whether the converse is always true is not established. Now according to Theorem 3.1, since the system is nearly AP, we can expect the singular values to be separated into two clusters; one having high and approximately equal values and the other having relatively small values. There must always be a large separation of the singular values since otherwise the system cannot be nearly AP. Hence is always a partition as in (19) such that $\sigma_{\check{m}} \gg \sigma_{\check{m}+1}$ and the best way to truncate is according to (19) and this will give us the IIR filter which is closest to an AP.

If a transfer function is AP, then the numerator polynomial is a mirror image of the denominator polynomial. Therefore if an IIR filter is nearly AP then the numerator polynomial is nearly a mirror image of the denominator. Now since we are looking for AP polyphase components, we must force the nearly AP transfer function to be AP. Hence we choose the numerator polynomial as the mirror image of the denominator polynomial. Alternatively we could choose the denominator as the mirror image of the numerator. However we are not guaranteed that the mirror image of the numerator polynomial is stable, whereas it is guaranteed that the denominator polynomial is stable.

3.3 Frequency Error Bounds

We shall use the result in (23) to obtain bounds for the magnitude response error and the phase response error when the FIR prototype filter is approximated with an IIR filter with AP polyphase components. This will show that the phase response error of the IIR filter is negligible.

Let the prototype FIR filter designed as in section 2, have a polyphase decomposition

$$H(z) = 0.5[X_0(z^2) + z^{-1}X_1(z^2)] \tag{24}$$

where $H(z)$ is linear phase and $|X_0(e^{j2\omega})| = |X_1(e^{j2\omega})| \approx 1$. Let us approximate $H(z)$ with

$$\tilde{H}(z) = 0.5[A_0(z^2) + z^{-1}A_1(z^2)] \tag{25}$$

where $A_0(z)$ and $A_1(z)$ are AP.

Theorem 3.2 *The frequency and phase response errors of the approximation are bounded as*

$$\|H(e^{j\omega}) - \tilde{H}(e^{j\omega})\|_\infty \leq \frac{1}{2}\|X_0(e^{j2\omega}) - A_0(e^{j2\omega})\|_\infty + \frac{1}{2}\|X_1(e^{j2\omega}) - A_1(e^{j2\omega})\|_\infty \tag{26}$$

$$\|\phi(\omega) - \tilde{\phi}(\omega)\|_\infty \leq \|X_0(e^{j2\omega}) - A_0(e^{j2\omega})\|_\infty + \|X_1(e^{j2\omega}) - A_1(e^{j2\omega})\|_\infty \tag{27}$$

provided that each of the quantities on the r.h.s are $\ll 1$.

Proof:

The claim in (26) can be easily verified. Now due to (10) we have $|X_0(e^{j2\omega})| = |X_1(e^{j2\omega})|$. Therefore if the arguments of $X_0(e^{j2\omega})$ and $X_1(e^{j2\omega})$ are $\phi_0(\omega)$ and $\phi_1(\omega)$ respectively we have

$$H(e^{j\omega}) = 0.5|X_0(e^{j2\omega})|[e^{j\phi_0} + e^{-j\omega}e^{j\phi_1}] \tag{28}$$

$$= |X_0(e^{j2\omega})| \cos(\frac{\phi_0 - \phi_1 + \omega}{2})e^{(\phi_0+\phi_1-\omega)/2} \tag{29}$$

Hence the phase of $H(e^{j\omega})$ is given by

$$\phi(\omega) = \frac{\phi_0 + \phi_1 - \omega}{2} \tag{30}$$

similarly if the arguments of $A_0(e^{j2\omega})$ and $A_1(e^{j2\omega})$ are $\tilde{\phi}_0(\omega)$ and $\tilde{\phi}_1(\omega)$ respectively we have

$$\tilde{H}(e^{j\omega}) = \cos(\frac{\tilde{\phi}_0 - \tilde{\phi}_1 + \omega}{2})e^{(\tilde{\phi}_0+\tilde{\phi}_1-\omega)/2} \tag{31}$$

Hence the phase of $\tilde{H}(e^{j\omega})$ is given by

$$\tilde{\phi}(\omega) = \frac{\tilde{\phi}_0 + \tilde{\phi}_1 - \omega}{2} \tag{32}$$

$$\text{So} \quad |\phi(\omega) - \tilde{\phi}(\omega)| = 0.5|(\phi_0 - \tilde{\phi}_0) + (\phi_1 - \tilde{\phi}_1)| \tag{33}$$

$$\leq 0.5(|\phi_0 - \tilde{\phi}_0| + |\phi_1 - \tilde{\phi}_1)|) \tag{34}$$

$$\text{Now} \quad |\phi_0(\omega) - \tilde{\phi}_0(\omega)| = \left|\ln\left(\frac{X_0(e^{j2\omega})|A_0(e^{j2\omega})|}{A_0(e^{j2\omega})|X_0(e^{j2\omega})|}\right)\right| \tag{35}$$

$$= |\ln(1 + \delta_0(\omega))| \tag{36}$$

$$\approx |\delta_0(\omega)| \quad \text{when} \quad |\delta_0| \ll 1 \tag{37}$$

where

$$|\delta_0(\omega)| = \left|1 - \frac{X_0(e^{j2\omega})|A_0(e^{j2\omega})|}{A_0(e^{j2\omega})|X_0(e^{j2\omega})|}\right| \tag{38}$$

$$\leq \left|\frac{A_0(e^{j2\omega}) - X_0(e^{j2\omega})}{A_0(e^{j2\omega})}\right| + \left|\frac{|X_0(e^{j2\omega})| - |A_0(e^{j2\omega})|}{A_0(e^{j2\omega})}\right|$$

$$\leq 2|A_0(e^{j2\omega}) - X_0(e^{j2\omega})| \tag{39}$$

where we have used the fact that $|A_0(e^{j2\omega})| = 1$. Therefore, from (37) and(39)

$$|\phi_0(\omega) - \tilde{\phi}_0(\omega)| \leq 2|A_0(e^{j2\omega}) - X_0(e^{j2\omega})| \tag{40}$$

Similarly, we can show that

$$|\phi_1(\omega) - \tilde{\phi}_1(\omega)| \leq 2|A_1(e^{j2\omega}) - X_1(e^{j2\omega})| \tag{41}$$

Therefore, the claim follows from (34),(40) and (41)

Now the BA of $X_0(z)$ is the nearly AP transfer function $\check{X}_0(z)$. We choose $A_0(z)$ to be the AP filter such that its denominator is equal to that of $\check{X}_0(z)$. Then it turns out that the numerator coefficients of $A_0(z)$ and $\check{X}_0(z)$ are very close due to the approximately AP nature of $\check{X}_0(z)$. In Appendix A we show that the frequency response error between $A_0(z)$ and $\check{X}_0(z)$ is small provided that their corresponding numerator coefficients are approximately equal and that $\check{X}_0(z)$ has a good stability margin. Therefore

$$||\check{X}_0(z) - A_0(z)||_\infty \leq \epsilon_0 \tag{42}$$

$$||\check{X}_1(z) - A_1(z)||_\infty \leq \epsilon_1 \tag{43}$$

where ϵ_0 and ϵ_1 are small. From (23), (42), (43) and Theorem 3.2, it follows that

$$\|H(e^{j\omega}) - \tilde{H}(e^{j\omega})\|_\infty \leq 2\sum_{i=0}^{1}\sum_{j=\tilde{m}_i+1}^{m_i}\sigma_{ij} + \epsilon_0 + \epsilon_1 \tag{44}$$

$$\|\phi(\omega) - \tilde{\phi}(\omega)\|_\infty \leq 2\sum_{i=0}^{1}\sum_{j=\tilde{m}_i+1}^{m_i}\sigma_{ij} + \epsilon_0 + \epsilon_1 \tag{45}$$

where σ_{ij} are the discarded singular values of $X_i(z)$ respectively.

From these results, we conclude that with the BA technique, the magnitude response error and the phase response error of the IIR approximation are small since the discarded Hankel singular values of the two approximately AP polyphase components of the FIR prototype filter are small.

To illustrate the application of the BR procedure to obtain an IIR allpass filter from a nearly allpass FIR filter we present an example.

Example 4.1: We start with an approximately allpass FIR filter given by

$$H(z) = -0.0076 - 0.0054z^{-1} + 0.1769z^{-2} + 0.9688z^{-3} - 0.1694z^{-4} + 0.0377z^{-5}$$

This filter is $2E_0(z)$ (i.e., the 0^{th} polyphase component with the magnitude normalized) of the Johnston 12A filter tabulated in [4]. The canonical s.s realization of this FIR filter is given in Table I.

Filter	A					B	C	D
FIR	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	-0.005421	-0.007619
	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.176940	
	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.968779	
	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000	-0.169391	
	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.037713	
IIR	-0.092316	-0.617899	0.327138			-0.709002	0.709002	-0.007619
	0.617899	0.565276	0.369593			-0.402579	-0.402579	
	-0.327138	0.369593	-0.648984			-0.578959	-0.578959	

Table I: State space realizations of the FIR and IIR filters in Example 4.1.

A balancing transformation that diagonalizes the gramians is

$$T = \begin{bmatrix} -0.7090 & -0.4026 & -0.5790 & -0.2219 & 0.2093 \\ 0.1248 & -0.8796 & 0.4589 & -0.1408 & 0.9943 \\ 0.6822 & -0.2504 & -0.6637 & 5.2256 & -0.9103 \\ -0.1248 & 0.0350 & 0.1154 & 28.1241 & -30.6132 \\ 0.0267 & -0.0152 & -0.0218 & -7.2917 & -118.7535 \end{bmatrix}$$

The diagonal elements of the diagonalized gramians are

$$diag\Sigma = \begin{bmatrix} 1.000036 & 1.000028 & 0.999998 & 0.001147 & 0.000066 \end{bmatrix}$$

It is seen that the first three elements are close to 1 and the rest is much smaller. Hence the order of the IIR filter is 3. Next the balanced system is found and truncated to obtain the 3^{rd} order approximately allpass IIR filter in Table I. The transfer function of this IIR filter is

$$\tilde{H}(z) = \frac{-0.007619 - 0.006762z^{-1} + 0.176036z^{-2} + 1.000014z^{-3}}{1.000000 + 0.176024z^{-1} - 0.006907z^{-2} - 0.008608z^{-3}}$$

which is approximately allpass. Hence according to the proposed method by choosing the numerator as the mirror image of the denominator the allpass filter is

$$A(z) = \frac{-0.008608 - 0.006907z^{-1} + 0.176024z^{-2} + 1.000000z^{-3}}{1.000000 + 0.176024z^{-1} - 0.006907z^{-2} - 0.008608z^{-3}}$$

The frequency response error bound for the BR is

$$\|H(e^{j\omega}) - \tilde{H}(e^{j\omega})\|_\infty \leq 0.004852$$

Hence the phase error bound using (40) is

$$\|\phi_H(\omega) - \phi_A(\omega)\|_\infty \leq (0.004852 + \epsilon)\ rad$$

Where ϵ is a small quantity. The actual phase error between $H(e^{j\omega})$ and $\tilde{H}(e^{j\omega})$ is 0.002130 rad and that between $H(e^{j\omega})$ and $A(e^{j\omega})$ is

$$\|\phi_H(\omega) - \phi_A(\omega)\|_\infty = 0.003138\ rad$$

This shows that the error bound gives a good estimate of the error that can be expected.

The balanced approximation is not the optimal lower order approximation, but we choose it because of its computational advantage. Also the question arises as to whether the optimal AP approximation to a nearly AP IIR filter is obtained by choosing the numerator polynomial to be the mirror image of the denominator polynomial. The only reason for this choice is because the BA is guaranteed to provide a stable transfer function, the denominator polynomial must be stable. Another choice is to choose the denominator polynomial to be the mirror image of the numerator. However in this instance we cannot guarantee that the resulting filter will be stable. Due to these two facts we cannot be sure whether this gives the AP filter which is closest to the original. Hence we optimize the coefficients of the filter to obtain an optimal filter.

4 Optimization of the Filter Parameters

We shall now use the filter parameters obtained by the application of the BR procedure as initial values and find the optimum parameters for the filter. In the design of IIR digital filters the most popular method is the Fletcher-Powell optimization procedure [23]. This algorithm has several advantages. Only first derivative information of the objective function is required, which in this case is readily obtained as we shall show. It has been used with success by many people [24][25], demonstrating its good convergence properties. A programmed version of the algorithm is available in the MATLAB optimization toolbox. This program allows bounds to be defined on the filter parameters, which is useful to ensure that the filter remains stable.

We shall not discuss the optimization procedure, but merely state that the rate of convergence of the method depends on the accuracy of the first derivative information of the objective function. Hence, although it is possible to compute the first derivatives numerically, we shall use analytical expressions.

4.1 The Form of the Filter Transfer Function

The filter transfer function takes the form in equation (31). The optimization is done for the filter $H_0(z)$ and not for the AP polyphase components. The order of these two AP filters are the same as the order of the AP filters obtained by applying the BR procedure to the polyphase components. Let each $A_i(z)$ obtained by applying the BR procedure have $(m_i + 2n_i)$ poles of which $2n_i$ are complex pairs. Then the AP filters take the form.

$$A_i(z) = \prod_{k=1}^{n_i} \frac{(r_{i,k}e^{j\phi_{i,k}} + z^{-1})(r_{i,k}e^{-j\phi_{i,k}} + z^{-1})}{(1 + r_{i,k}e^{-j\phi_{i,k}}z^{-1})(1 + r_{i,k}e^{j\phi_{i,k}}z^{-1})} \prod_{k=n_i+1}^{m_i} \left(\frac{r_{i,k} + z^{-1}}{1 + r_{i,k}z^{-1}}\right), \quad i = 0, 1 \tag{46}$$

We choose this cascade form for the polyphase components due to several reasons. Firstly, stability of a cascade filter is readily tested. This is important since during the process of optimization, we have to ensure that the filter is stable. Secondly, errors due to quantization and finite wordlength size are much less relative to other forms [26]. Finally, the magnitude response and the group delay of $H_0(z)$ takes a particularly simple functional form, permitting easy calculation of the first derivatives.

4.2 The Objective Function

Our objective in the optimization procedure is to minimize the magnitude response error and the group delay distortion. Hence our objective function is

$$f(\mathbf{x}) = \gamma \oint (\tilde{M}_0(e^{j\omega}) - M_0(e^{j\omega}))^2 d\omega \; + \; (1-\gamma) \oint (\tilde{\tau} - \tau)^2 d\omega \quad 0 < \gamma < 1 \tag{47}$$

where $\mathbf{x} = (r_{01}, r_{02}, \ldots, r_{0m_0}, r_{11}, r_{12}, \ldots, r_{1m_1}, \phi_{01}, \phi_{02}, \ldots, \phi_{0n_0}, \phi_{11}, \phi_{12}, \ldots, \phi_{1n_1})$, $M_0(e^{j\omega})$ and $\tilde{M}_0(e^{j\omega})$ are differentiable magnitude response functions of the prototype FIR filter and the IIR filter respectively and τ, $\tilde{\tau}$ are the respective group delays. A differentiable magnitude response function of $\tilde{H}_0(e^{j\omega})$ is readily obtained from (31) as

$$\tilde{M}_0(e^{j\omega}) = \cos(\alpha_1 - \alpha_0 - \omega/2) \tag{48}$$

where

$$\alpha_i = \sum_{j=0}^{1}\sum_{k=1}^{n_i} \arctan\left(\frac{r_{i,k}\sin(2\omega+(-1)^j\phi_{i,k})}{1+r_{i,k}\cos(2\omega+(-1)^j\phi_{i,k})}\right) + \sum_{k=n_i+1}^{m_i} \arctan\left(\frac{r_{i,k}\sin(2\omega)}{1+r_{i,k}\cos(2\omega)}\right) - (m_i+n_i)\omega \tag{49}$$

and the group delay is given by

$$\tilde{\tau} = 0.5 + \sum_{i=0}^{1}\left[\sum_{k=1}^{n_i}\sum_{j=0}^{1} \frac{1-r_{i,k}^2}{1+2r_{i,k}\cos(2\omega+(-1)^j\phi_{i,k})+r_{i,k}^2} + \sum_{k=n_i+1}^{m_i} \frac{1-r_{i,k}^2}{1+2r_{i,k}\cos(2\omega)+r_{i,k}^2}\right] \tag{50}$$

For completeness we give the partial derivatives of $f(\mathbf{x})$. The partial derivatives of the objective function $f(\mathbf{x})$ w.r.t x, where x is an element of $\mathbf{x}$ will be

$$\frac{\partial f(\mathbf{x})}{\partial x} = 2\gamma \oint (\tilde{M}_0(z) - M_0(z))\frac{\partial \tilde{M}_0(z)}{\partial x}dz \; + \; 2(1-\gamma)\oint (\tilde{\tau}-\tau)\frac{\partial \tilde{\tau}}{\partial x}dz \tag{51}$$

where if $i = 0, 1$ and $j = 1, 2, \ldots, n_i$

$$\frac{\partial \tilde{M}_0(z)}{\partial \phi_{ij}} = (-1)^i \sin(\alpha_1 - \alpha_0 - \omega/2)\sum_{k=0}^{1}(-1)^k r_{ij}\left(\frac{\cos(2\omega+(-1)^k\phi_{ij})+r_{ij}}{1+2r_{ij}\cos(2\omega+(-1)^k\phi_{ij})+r_{ij}^2}\right) \tag{52}$$

$$\frac{\partial \tilde{M}_0(z)}{\partial r_{ij}} = \sin(\alpha_1 - \alpha_0 - \omega/2)\sum_{k=0}^{1}\frac{\sin(2\omega+(-1)^k\phi_{ij})}{1+2r_{ij}\cos(2\omega+(-1)^k\phi_{ij})+r_{ij}^2} \tag{53}$$

$$\frac{\partial \tilde{\tau}}{\partial r_{ij}} = -2\sum_{k=0}^{1}\frac{\cos(2\omega+(-1)^k\phi_{ij})+2r_{ij}+r_{ij}^2\cos(2\omega+(-1)^k\phi_{ij})}{(1+2r_{ij}\cos(2\omega+(-1)^k\phi_{ij})+r_{ij}^2)^2} \tag{54}$$

$$\frac{\partial \tilde{\tau}}{\partial \phi_{ij}} = -2\sum_{k=0}^{1}\frac{(-1)^k r_{ij}(1-r_{ij}^2)\sin(2\omega+(-1)^k\phi_{ij})}{(1+2r_{ij}\cos(2\omega+(-1)^k\phi_{ij})+r_{ij}^2)^2} \tag{55}$$

and if $i = 0, 1$ and $j = n_i+1, n_i+2, \ldots, m_i$

$$\frac{\partial \tilde{M}_0(z)}{\partial r_{ij}} = \frac{\sin(\alpha_1 - \alpha_0 - \omega/2)\sin 2\omega}{1+2r_{ij}\cos 2\omega + r_{ij}^2} \tag{56}$$

$$\frac{\partial \tilde{\tau}}{\partial r_{ij}} = \frac{-2(\cos 2\omega + 2r_{ij} + r_{ij}^2\cos 2\omega)}{(1+2r_{ij}\cos 2\omega + r_{ij}^2)^2} \tag{57}$$

In the optimization procedure we minimize $f(\mathbf{x})$ subject to $|r_{ij}| < 1$.

5 Design Examples

In this section, we demonstrate the application of the proposed method to design an IIR filter bank with each analysis filter having a stop band edge $\omega_s = 0.586$ and a stop-band attenuation of 65dB. Results of the design of two other filters which have the same ω_s, but different stopband energies are also given. A comparison is made with comparable Type I and Type II FIR systems.

Example 5.1: The design of FIR filters with the desired properties referred to in this paper have been presented earlier by Johnston [4]. We choose a filter which satisfies the above specifications from the appendix of [4]. This turns out to be a filter of length 64, which is referred to as the $64D$ filter.

The magnitude response of this filter is shown in Fig. 3 and those of the two polyphase components is shown in Fig. 4. This shows that the polyphase components are approximately AP.

We now find the balanced realization for each polyphase component. The largest twenty singular values of the realization are tabulated in Table II. It can be clearly seen that each polyphase component has a dominant subsystem which is AP. For the 0^{th} polyphase component, the order of the AP subsystem is 16, while that for the 1^{st} polyphase component is 15. However we must mention that these singular values are not exactly equal. There are differences of the order of 10^{-5} in these singular values. Nevertheless, these are very closely clustered together and we can expect it to be nearly AP. Now we truncate the two polyphase systems and find the IIR approximations. The two denominator polynomials of the truncated systems, $\check{d}_0(z)$ and $\check{d}_1(z)$ are shown in Table III. We choose the numerator polynomial of each polyphase component to be the mirror image of the respective denominator polynomials.

m	$\sigma_{0,m}$	$\sigma_{1,m}$
1	0.5000	0.5000
2	0.5000	0.5000
3	0.5000	0.5000
4	0.5000	0.5000
5	0.5000	0.5000
6	0.5000	0.5000
7	0.5000	0.5000
8	0.5000	0.5000
9	0.5000	0.5000
10	0.5000	0.5000
11	0.5000	0.5000
12	0.5000	0.5000
13	0.5000	0.5000
14	0.5000	0.5000
15	0.5000	0.5000
16	0.5000	0.0001
17	0.0001	0.0001
18	0.0000	0.0000
19	0.0000	0.0000
20	0.0000	0.0000

Table II: Largest 20 Hankel singular values for the two polyphase components in Example 5.1.

The magnitude response and the group delay of the IIR filter $\check{H}_0(z) = 0.5[\check{A}_0(z^2) + z^{-1}\check{A}_1(z^2)]$ is shown in Fig. 5 where the denominator polynomials of the AP polyphase components are given in Table III. One can see that the approximation error is very small. The theoretical error bound for the phase response with the BR procedure given in (45), neglecting ϵ_0 and ϵ_1, is 7.4×10^{-4}, while the actual phase error for the AP approximation is 4.3×10^{-4}. This indicates that the theoretical error bound

for the BR procedure gives a good estimate of the actual error. The computer time taken for the BR procedure is approximately 0.64 seconds and 1.1×10^6 floating point operations (FLOPS) on a DEC 5000 workstation using the MATLAB control system toolbox.

Now with the parameters of $\check{H}_0(z)$ as the initial values we find the filter $\tilde{H}_0(z)$ which is optimal in the sense that (47) is minimized. The denominator polynomials, $\tilde{d}_0(z)$ and $\tilde{d}_1(z)$ of the polyphase components of $\tilde{H}_0(z)$ are given in table II and the magnitude response and group delay are shown in Fig. 6. In this example the improvement in the group delay distortion is only marginal indicating that the BR procedure by itself yields nearly optimal filters. The computer time for the optimization requires about 5 minutes. We could achieve the same results with a direct optimization without using the BR technique, but this requires computer time of 17 minutes!

Example 5.2: We could also use the optimization procedure to reduce the group delay distortion at the expense of increased stopband energy. If we use the $48D$ FIR filter as the prototype, we obtain an IIR approximation with stop band attenuation greater than 50 dB and and maximum group delay distortion ± 0.06 samples. With optimization we can reduce the maximum groupdelay distortion to ± 0.0125 samples if a stop band attenuation of 42.5 dB is satisfactory. The denominator coefficients of this filter are given in Table B-I of the appendix.

Example 5.3: The denominator coefficients of a design using a lower order FIR (32D Johnston Filter) is shown in Table B-II of Appendix B. In this case since the order of the IIR filter is lower, the stopband attenuation is also low and the group delay distortion is also somewhat higher than the higher order designs. In this case no appreciable improvement in the group delay distortion could be obtained at the expense of stopband attenuation.

The maximum stopband attenuation that can be achieved is the stopband attenuation of the prototype filter. This is because we minimize the frequency response error of the IIR with respect to the FIR prototype. If one can tolerate more group delay distortion but needs higher stopband attenuation a FIR prototype filter has to be designed with the required stopband attenuation while allowing for a larger AMD. This can be achieved by choosing a larger weighting factor α in (8).

n	$d_0(n)$	$\tilde{d}_0(n)$	$d_1(n)$	$\tilde{d}_1(n)$
0	1.00000000000000	1.00000000000000	1.00000000000000	1.00000000000000
1	0.24510479134270	0.24510472044679	-0.24510331217525	-0.24510335220548
2	-0.08540446106043	-0.08540104480996	0.14547975580225	0.14548313728944
3	0.04372217148846	0.04372495061969	-0.10031250245546	-0.10031129317841
4	-0.02473518071621	-0.02473108389726	0.07246314200969	0.07246687835227
5	0.01413313848422	0.01412938906084	-0.05288462490705	-0.05289001097297
6	-0.00769194123910	-0.00768883641505	0.03829127599374	0.03829712692391
7	0.00369791714860	0.00369704301735	-0.02719069757361	-0.02719472555477
8	-0.00128353807212	-0.00127508069423	0.01876589218446	0.01877657358317
9	-0.00009405863941	-0.00013315486101	-0.01245789864033	-0.01250200323171
10	0.00068111709418	0.00065913509141	0.00794962259068	0.00794917507237
11	-0.00089067178133	-0.00088585259755	-0.00477740314111	-0.00477542934039
12	0.00082093773822	0.00082071701203	0.00267592696004	0.00267324786445
13	-0.00062700877872	-0.00062822205023	-0.00136646062031	-0.00136132437679
14	0.00039360412931	0.00039431963754	0.00062174841299	0.00060977797019
15	-0.00019341002057	-0.00019017326905	-0.00023563382949	-0.00018118324745
16	0.00008049208892	0.00011897794914	-	-

Table III: Denominator polynomial coefficients of the two polyphase components in Example 5.1.

5.1 Effects of Coefficient Quantization

Since a digital filter is always implemented with a finite wordlength (FWL), it is important to determine the performance of the filter when the designed coefficients are quantized. We assume the filter is to be implemented in a floating point processor with a finite number of bits in the mantissa. The following filters were implemented in FWL:

- The 64D Johnston Type I filter.
- The 64 length Type II filter presented in [9] implemented in a '1-multiplier' lattice structure [3].
- The IIR filters designed using the 32D, 48D and 64D Johnston filters as the prototypes.

A 12-bit mantissa was used on the implementation of all three filters. Fig. 7 shows the overall magnitude response $|T(e^{j\omega})|$ for the Type I and Type II filter. For the IIR filter $|T(e^{j\omega})|$ is perfectly flat and for the Type I filter it is almost flat. One can see that the Type II system is far from a PR system under coefficient quantization. The Type I system seems far better than the Type II system. For Type I and Type II systems the groupdelay is perfectly flat, since the linear phase property is guaranteed under any level of coefficient quantization. For the IIR filter banks the groupdelay distortion increases marginally. Fig. 8 shows the magnitude response of the Type II, Type I and IIR lowpass filters. It is clearly seen that the Type II system is extremely sensitive to coefficient quantization effects. The IIR and Type I filters have very low sensitivity. As seen in Fig. 8, even for a high order filter such as the 64D filter, there is no appreciable change in the magnitude response. Results are tabulated in Table IV.

The IIR filter designed in [10] with approximate linear phase characteristics in the pass band was

also implemented in FWL. This filter has a stop band attenuation of 71.2 dB but has a groupdelay distortion of about +16 samples. When implemented with a 12-bit mantissa the stopband attenuation dropped marginally to 69.4 dB.

5.2 Effects of Roundoff Noise

To investigate effects of roundoff noise, the filter banks were implemented with FWL and infinite wordlength (IWL). FWL implementation was with 12-bits in the mantissa and using double precision for the internal registers (i.e., intermediate values are stored as a floating point number with 24-bits in the mantissa). Due to practical constraints, in the IWL implementation the number of bits in the mantissa was 26. The input signal was white noise with a flat spectrum and random phase. The signal was passed through the filterbank and the signal was reconstructed at the output. The signal-to-noise ratio (SNR) at the output was calculated using

$$SNR = \sum_{n=0}^{N} \frac{\hat{x}(n)^2}{(x(n) - \hat{x}(n))^2} \tag{58}$$

The results are tabulated in Table IV.

For the IIR filter in [10] the SNR was only about 5 dB (for IWL and FWL). However when a fairly smooth signal (band limited to the lower end of the frequency spectrum) was used the SNR increased to 45 dB. Hence the very poor SNR can be attributed to the very large groupdelay distortion. Hence for this filter bank too, a postprocessing AP equalizer network is necessary.

5.3 Comparison with FIR filters

An AP filter of order p with real coefficients can be implemented with p multipliers and $2p$ adders [27]. However if we choose the structure in Fig. 2, each polyphase component is computing at half rate. Hence each polyphase component needs only $p/2$ multiplications per unit time (MPU's) and p additions per unit time (APU's), where p is the order of that polyphase component. Hence our analysis bank in example 5.1 needs only $(16 + 15)/2 = 15.5$ MPU's and $(31 * 2 + 2)/2 = 32$ APU's. Table IV gives a comparison of three different IIR filters with FIR filters. The data for the Type II filter is based on [11]. The three IIR filters labeled (1), (2) and (3) correspond to designs based on the prototype FIR filters 64D, 48D and 32D respectively (examples 5.1, 5.2 and 5.3).

Feature		Type I FIR	Type II FIR	IIR (1)	IIR (2)	IIR (3)
Distortions in Filter Bank		ALD canceled AMD minimized PHD eliminated	ALD canceled AMD eliminated PHD eliminated	ALD canceled AMD eliminated PHD minimized		
No of MPU's for Analysis bank		32	17	15.5	11.5	7.5
No of APU's for Analysis Bank		32	49	32	24	16
Average group delay (samples)		63	63	63	47	31
Stop Band Atten.	- IWL	65 dB	42 dB	65 dB	53.1 dB	37.5 dB
	- FWL	68 dB	06 dB	66.5 dB	53 dB	36.9 dB
AMD	- IWL	0.002 dB	0 dB	0 dB	0 dB	0 dB
	- FWL	0.006 dB	3.320 dB	0 dB	0 dB	0 dB
Group delay distortion (samples)	- IWL	0	0	± 0.0125	± 0.0400	± 0.0231
	- FWL	0	0	± 0.0309	± 0.0400	± 0.0579
SNR	- IWL	144 dB	179 dB	143 dB	135 dB	117 dB
	- FWL	125 dB	21 dB	125 dB	124 dB	115 dB

Table IV: Comparison of three different IIR filter designs with FIR filters.

Based on the information in the Table IV the IIR is better than the Type I and Type II filters in every respect excepting the group delay distortion. When compared with the Type I filters, the IIR filter would be much better when efficient implementation is necessary since the MPU count is very much less for the latter while the same stopband attenuation is achieved. It is obvious that the Type II filters, to be of any practical use, have to be implemented with the 2-multiplier lattice structure. Then the Type II filter will have the same MPU count as for the Type I filter. Hence again the IIR filter is much more efficient than the Type II filter. The price paid for the better efficiency is the small group delay distortion. However since the group delay is very small, for most applications this is bound to be acceptable.

6 Conclusion

We have presented an algorithm to design IIR filter banks based on FIR prototypes. The analysis filters of the IIR filter bank have approximately linear-phase. It was shown that the Type I FIR filter family was the most suitable prototype filter. Although it is possible to obtain the IIR filter bank with direct optimization, the design becomes computationally more efficient when the BR scheme is

used to obtain initial values for the filter coefficients. Furthermore the partial error bounds for the BR scheme, gives a good estimate of the phase error that can be achieved. Design examples demonstrated the application of the algorithm and indicated the computational advantage of the IIR filter bank compared to FIR designs. In most cases the the BR procedure gives good IIR filters needing little or no optimization. However, optimization could be used to obtain lower group delay distortion at the expense of lower stopband attenuation. The group delay distortion of the filter bank is so small that for most applications an IIR implementation seems to be better than an FIR implementation due to the computational advantage.

A further simplification of the method is possible due to the work in [28], where, an algorithm to obtain a reduced order IIR filter from a high order FIR filter, without explicitly constructing an interim balanced realization, has been presented. This method is much simpler than the conventional method of BR.

As indicated by the simulations with FWL, efficiently implemented Type II systems lack robustness under coefficient quantization. Furthermore the stopband characteristics of these filters degrade under coefficient quantization and hence the 2-multiplier lattice structure should be used to implement Type II filters. Hence robust Type II filter banks are not as efficient as IIR filter banks. The IIR filters are not sensitive to coefficient quantization effects mainly due to the structurally robust allpass polyphase structure. The Type I filters too exhibit good robustness under coefficient quantization.

In summary the IIR filter banks have all the desired properties, viz., good stopband attenuation, low computational complexity, linear phase (approximately), low reconstruction error and low sensitivity.

An alternative approach to design the proposed IIR filter banks is to use the eigenfilter method [10] to approximate the phase of the AP polyphase components with that of the prototype filter polyphase components. This method yields IIR filter banks which are nearly identical to those obtained with the proposed method. However this method requires more computer time and more number of operations than the proposed method.

Further research problems would be to investigate the improvement in performance of the filter banks when implemented in δ-operator formulation [29][30] and the extension of the proposed method to the 2-dimensional case using the BR algorithm presented in [31].

Acknowledgement

We are grateful to the anonymous reviewers for the usefull suggestions which helped in improving the paper and to Dr. Truong Nguyen for providing [10].

Appendix A

Let $T(z)$ be a nearly AP transfer function with

$$T(z) = \prod_{k=1}^{n} \frac{(r_k e^{j\phi_k} + z^{-1})(r_k e^{-j\phi_k} + z^{-1})}{(1 + \tilde{r}_k e^{-j\tilde{\phi}_k} z^{-1})(1 + \tilde{r}_k e^{j\tilde{\phi}_k} z^{-1})} \prod_{k=n+1}^{m} \left(\frac{r_k + z^{-1}}{1 + \tilde{r}_k z^{-1}} \right) \tag{59}$$

where we assume that $\sup_k |\tilde{r_k} - r_k| \leq \Delta r$ and $\sup_k |\tilde{\phi}_k - \phi_k| \leq \Delta\phi$. This is true since $T(z)$ is nearly AP. Note that $\tilde{T}(z) = T(z)|_{p=\tilde{p}}$ is AP, where $p = (p_1, p_2, \ldots, p_m)$ and $\tilde{p} = (\tilde{p}_1, \tilde{p}_2, \ldots, \tilde{p}_m)$, $p_k = (r_k, \phi_k)$, $\tilde{p}_k = (\tilde{r}_k, \tilde{\phi}_k)$, The group delay is given by

$$\tau = 0.5 \left[\sum_{k=1}^{n} \sum_{j=0}^{1} \left(\frac{1 - \tilde{r}_k^2}{1 + 2\tilde{r}_k \cos(\omega + (-1)^j \tilde{\phi}_k) + \tilde{r}_k^2} + \frac{1 - r_k^2}{1 + 2r_k \cos(\omega + (-1)^j \phi_k) + r_k^2} \right) + \right. \tag{60}$$

$$\left. \sum_{k=n+1}^{m} \left(\frac{1 - \tilde{r}_k^2}{1 + 2\tilde{r}_k \cos(\omega) + \tilde{r}_k^2} + \frac{1 - r_k^2}{1 + 2r_k \cos(\omega) + r_k^2} \right) \right] \tag{61}$$

Hence we can write

$$\Delta\tau \approx \sum_{k=1}^{m} \left. \frac{\partial\tau}{\partial r_k} \right|_{p=\tilde{p}} \Delta r_k + \sum_{k=1}^{n} \left. \frac{\partial\tau}{\partial\phi_k} \right|_{p=\tilde{p}} \Delta\phi_k \tag{62}$$

where the higher order terms have been neglected. Therefore

$$|\Delta\tau| \leq \sum_{k=1}^{m} \left| \left. \frac{\partial\tau}{\partial r_k} \right|_{p=\tilde{p}} \right| |\Delta r_k| + \sum_{k=1}^{n} \left| \left. \frac{\partial\tau}{\partial\phi_k} \right|_{p=\tilde{p}} \right| |\Delta\phi_k| \tag{63}$$

Now
$$\frac{\partial\tau}{\partial r_k} = -\sum_{j=0}^{1} \frac{r_k^2 \cos(\omega + (-1)^j \phi_k) + 2r_k + \cos(\omega + (-1)^j \phi_k)}{[1 + 2r_k \cos(\omega + (-1)^j \phi_k) + r_k^2]^2} \tag{64}$$

But if
$$C(\alpha) = \frac{r^2 \cos\alpha + 2r + \cos\alpha}{1 + 2r\cos\alpha + r^2} \tag{65}$$

then
$$[C(\alpha)]_{max} = C(0) = 1 \tag{66}$$

and
$$[C(\alpha)]_{min} = C(\pi) = -1 \tag{67}$$

So $|C(\alpha)| \leq 1$. Hence

$$\left| \frac{\partial\tau}{\partial r_k} \right| \leq \sum_{j=0}^{1} \frac{1}{1 + 2r_k \cos(\omega + (-1)^j \phi_k) + r_k^2} \tag{68}$$

Next consider

$$\frac{\partial\tau}{\partial\phi_k} = (1 - r_k^2) \sum_{j=0}^{1} \frac{(-1)^{j+1} r_k \sin(\omega + (-1)^j \phi_k)}{[1 + 2r_k \cos(\omega + (-1)^j \phi_k) + r_k^2]^2}$$

But if
$$D(\alpha) = \frac{\sin\alpha}{1 + 2r\cos\alpha + r^2} \tag{69}$$

then
$$[D(\alpha)]_{max} = D(\alpha)|_{\alpha = \sin^{-1}\left(\frac{1-r^2}{1+r^2}\right)} = \frac{1}{1 - r^2} \tag{70}$$

and
$$[D(\alpha)]_{min} = D(\alpha)|_{\alpha=\sin^{-1}\left(-\frac{1-r^2}{1+r^2}\right)} = -\frac{1}{1-r^2} \tag{71}$$

Hence
$$\left|\frac{\partial\tau}{\partial\phi_k}\right| \leq r_k \sum_{j=0}^{1} \frac{1}{1+2r_k\cos(\omega+(-1)^j\phi_k)+r_k^2} \tag{72}$$

Therefore

$$|\Delta\tau| \leq \sum_{k=1}^{n}\sum_{j=0}^{1} \frac{|\Delta r_k| + \tilde{r}_k|\Delta\phi_k|}{1+2\tilde{r}_k\cos(\omega+(-1)^j\tilde{\phi}_k)+\tilde{r}_k^2} + \sum_{k=n+1}^{m} \frac{|\Delta\tilde{r}_k|}{1+2\tilde{r}_k\cos\omega+\tilde{r}_k^2} \tag{73}$$

$$\leq \frac{\Delta r + \tilde{R}\Delta\phi}{1-\tilde{R}^2}\sum_{k=1}^{n}\sum_{j=0}^{1} \frac{1-\tilde{r}_k^2}{1+2\tilde{r}_k\cos(\omega+(-1)^j\tilde{\phi}_k)+\tilde{r}_k^2} + \frac{\Delta r}{1-\tilde{R}^2}\sum_{k=n+1}^{m} \frac{1-\tilde{r}_k^2}{1+2\tilde{r}_k\cos\omega+\tilde{r}_k^2} \tag{74}$$

$$\leq \frac{\Delta r + \tilde{R}\Delta\phi}{1-\tilde{R}^2}\left[\sum_{k=1}^{n}\sum_{j=0}^{1} \frac{1-\tilde{r}_k^2}{1+2\tilde{r}_k\cos(\omega+(-1)^j\tilde{\phi}_k)+\tilde{r}_k^2} + \sum_{k=n+1}^{m} \frac{1-\tilde{r}_k^2}{1+2\tilde{r}_k\cos\omega+\tilde{r}_k^2}\right] \tag{75}$$

$$= \left(\frac{\Delta r + \tilde{R}\Delta\phi}{1-\tilde{R}^2}\right)\tilde{\tau} \tag{76}$$

where $\tilde{R} = \sup_k \tilde{r}_k$ and $\tilde{\tau} = \tau|_{p=\tilde{p}}$. Hence the difference in groupdelay between $T(z)$ and $\tilde{T}(z)$ is small provided that none of the poles of $T(z)$ are too close to the unit circle in the z-plane, i.e., $T(z)$ has good stability margin. We also know that the magnitudes of $T(z)$ and $\tilde{T}(z)$ are close. Hence we can conclude that $||T(e^{j\omega}) - \tilde{T}(e^{j\omega})||_\infty$ is small.

Appendix B

n	$d_0(n)$	$d_1(n)$
0	1.0000	1.0000
1	0.2437	-0.2437
2	-0.0829	0.1423
3	0.0407	-0.0956
4	-0.0213	0.0663
5	0.0106	-0.0456
6	-0.0042	0.0303
7	0.0005	-0.0190
8	0.0012	0.0109
9	-0.0016	-0.0057
10	0.0009	0.0030
11	0.0002	-0.0022
12	0.0013	-

Table B-I

n	$d_0(n)$	$d_1(n)$
0	1.0000	1.0000
1	0.2409	-0.2408
2	-0.0774	0.1354
3	0.0335	-0.0847
4	-0.0134	0.0523
5	0.0032	-0.0301
6	0.0012	0.0155
7	-0.0016	-0.0074
8	0.0043	-

Table B-II

References

[1] P.P. Vaidyanathan, *Multirate Systems and Filter Banks.* Englewood Cliffs, NJ: Prentice Hall, 1993.

[2] P.P. Vaidyanathan, " Multirate Digital Filters, Filter Banks, Polyphase Networks and Applications: A Tutorial," *Proceedings of the IEEE*, vol. 78, No 1, January 1990.

[3] Z. Doganata and P.P. Vaidyanathan, "On one-multiplier implementations of FIR lattice structures," *IEEE Trans. Circuits Syst.*, vol. 34, pp 1608-1609, Dec. 1987.

[4] J.D. Johnston, "A filter family designed for use in quadrature mirror filter banks,"in *Proc. IEEE Int. Conf. ASSP*, Apr. 1980, pp 291-294.

[5] V.K. Jain and R.E. Crochiere, "Quadrature mirror filter design in the time domain," *IEEE Trans. Acoust. Speech Proc.*, vol. 32, pp 353-361, Apr. 1984.

[6] M.J.T. Smith and T.P. Barnwell III, "Exact reconstruction techniques for tree structured subband coders,"*IEEE Trans. Acoust. Speech Signal Proc.*, vol. ASSP-34, pp 434-441, June 1986.

[7] P.P. Vaidyanathan, "Theory and design of M-channel maximally decimated quadrature mirror filters with arbitrary M, having perfect reconstruction property," *IEEE Trans. Acoust. Speech Signal Proc.*, vol. 35, pp 476-492, Apr. 1987.

[8] P.P. Vaidyanathan and P.Q. Hoang, "Lattice structures for optimal design and robust implementation of two-channel perfect reconstruction QMF banks," *IEEE Trans. Acoust. Speech Signal Proc.*, vol. 36, pp 81-94, Jan. 1988.

[9] T.Q. Nguyen and P.P. Vaidyanathan, "Two-channel perfect reconstruction FIR QMF structures which yield linear phase FIR analysis and synthesis filters," *IEEE Trans. Acoust. Speech Signal Proc.*, vol. 37, pp 676-690, May 1989.

[10] T.Q. Nguyen, T.I. Laakso and R.D. Koilpillai, "Eigenfilter approach for the design of allpass filters approximating a given phase response," *IEEE Trans. Signal Proc.*, Sept. 1994.

[11] P.P. Vaidyanathan, S.K. Mitra and Y. Neuvo, "A new approach to the realization of low sensitivity IIR digital filters," *IEEE Trans. Acoust. Speech Signal Proc.*, vol. 34, pp 350-361, Apr. 1986.

[12] T.P. Barnwell III, "Sub-band coder design incorporating recursive quadrature filters and optimum ADPCM coders," *IEEE Trans. Acoust. Speech Signal Proc.*, vol. 30, pp 751-765, Oct. 1982.

[13] T.A. Ramstad, "IIR filterbank for subband coding of images," in *Proc. IEEE Int. Symp. on circuits and systems*, Espoo, Finland, June 1988, pp 827-830.

[14] S. Basu, C. Chiang, H. Choi, "Causal IIR Perfect Reconstruction Subband Coding", ISCAS 1993, pp 367-370.

[15] M.M. Ekanayake and K. Premaratne, "Two-Channel IIR QMF Banks with Approximately Linear-Phase Analysis and Synthesis Filters", presented at the 28^{th} Annual Asilomar Conference on Signals, Systems, and Computers, Nov. 1994.

[16] B.C. Moore, "Principal Component Analysis in Linear Systems : Controllability, Observability and Model Reduction", *IEEE Trans. Auto. Control,* vol. AC-26, No 1, pp17-32, Feb. 1981.

[17] R.H. Bartels and G.W. Stewart, "Solution of the matrix equation $AX + XB = C$," *Comm ACM,* vol. 15, pp 820-826, 1972.

[18] A.J. Laub, "Computation of Balancing Transformations," *Proc. JACC,* vol. 1, 1980, session FA8-E.

[19] C.V.K. Prabakara Rao and P. DeWilde, "On lossless transfer functions and orthogonal realizations," *IEEE Trans. Circuits and Systems,* vol. 34, pp 677-678, June 1987.

[20] Pro-MATLAB Users Guide, *MathWorks, Inc.*

[21] U.M. Al-Saggaf and G.F. Franklin, "An Error Bound for a Discrete Reduced Order Model of a Linear Multivariable System", *IEEE Trans. Auto. Control,* vol. AC-32, No. 9, pp. 815-819, Sept. 1987.

[22] K. Glover, "All optimal Hankel-norm approximations of linear multivariable systems and their L^∞-error bounds," *Int. J. Control,* vol. 39, No 6, pp 1115-1193,1984.

[23] R. Fletcher and M.J.D. Powell, "A rapidly convergent descent method for minimization," *Comp. J.,* vol. 6, pp 163-168, 1963.

[24] A.G. Deczky, "Synthesis of Recursive Digital Filters Using the Minimum p-Error Criterion," *IEEE Trans. Audio and Electroacoustics,* vol. AU-20, No 4, pp. 257-263, Oct. 1972.

[25] G.C. Bown, "Design and optimization of circuits by computer," *Proc. Inst. Elec. Eng.,* vol. 118, pp. 649-661, May 1971.

[26] J.B. Knowles and E.M. Olcayto, "Coefficient accuracy and digital filter response," *IEEE Trans. Circuit Theory,* vol. CT-15, pp 31-41, Mar. 1968.

[27] A.H. Gray, Jr., " Passive cascaded lattice digital filters," *IEEE Trans. Circuits Syst.,* vol. 27, pp 337-344, May 1980.

[28] B. Beliczynski, I. Kale and G.D. Cain, "Approximation of FIR by IIR Digital Filters: An Algorithm Based on Balanced Model Reduction," *IEEE Trans. Signal Proc.,* vol. 40, No 3, pp 532-542, March 1992.

[29] G.C. Goodwin, R.H. Middleton and V. Poor, "High-Speed Digital Signal Processing and Control," *Proceedings of the IEEE,* vol. 80, No 2, pp 240-259, Feb. 1992.

[30] K. Premaratne and E.I. Jury, "Tabular method for determining root distribution of delta operator formulated real polynomials," *IEEE Trans. Automatic Control,* vol. 39, pp 352-355, Feb. 1994.

[31] K. Premaratne, E.I. Jury and M. Mansour, "An Algorithm for the Model Reduction of 2-D Discrete Time Systems," *IEEE Trans. Circuits Syst.,* vol. 37, No 9, pp 1116-1132, Sept. 1990.

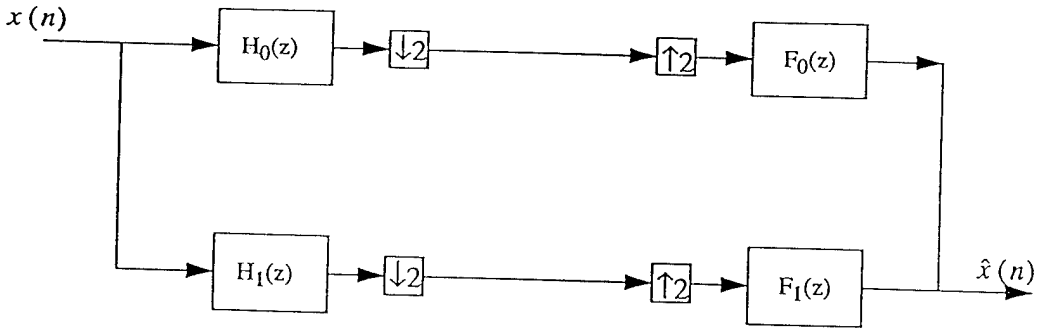


Figure 1: Two-channel QMF bank

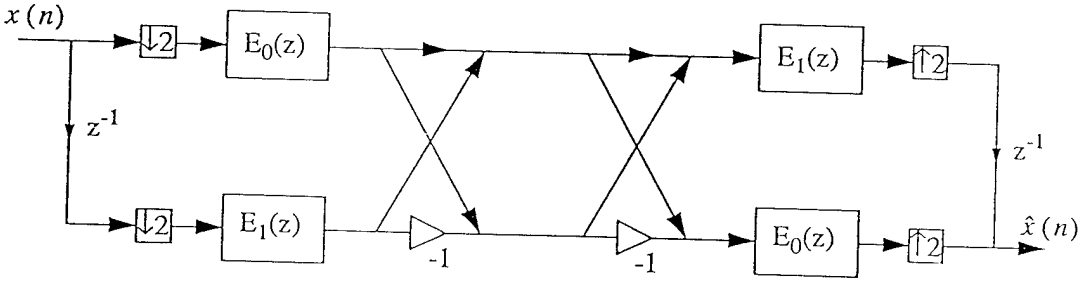


Figure 2: Polyphase implementation of the QMF bank

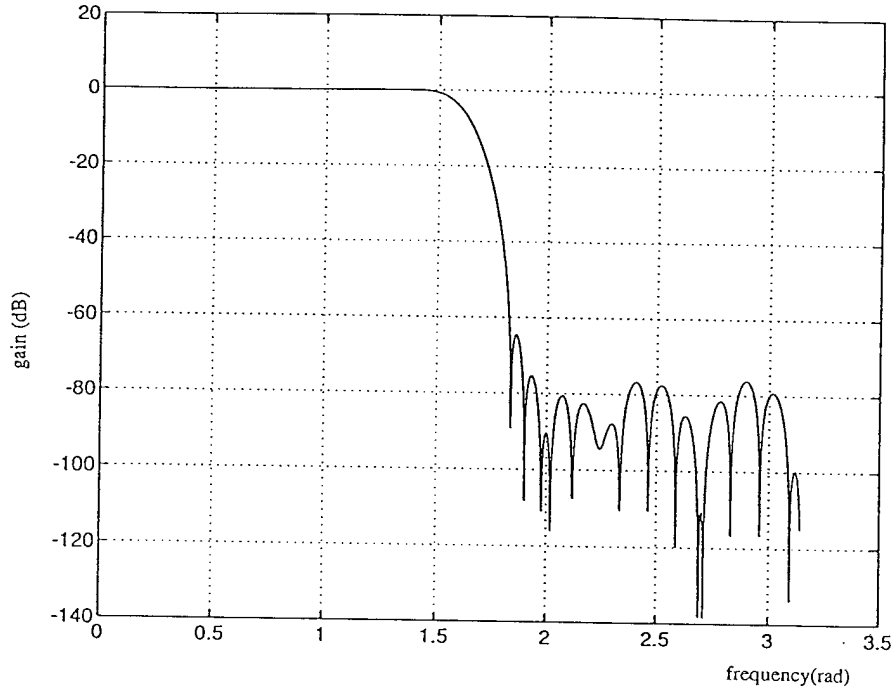


Figure 3: Magnitude response of 64D Johnston filter

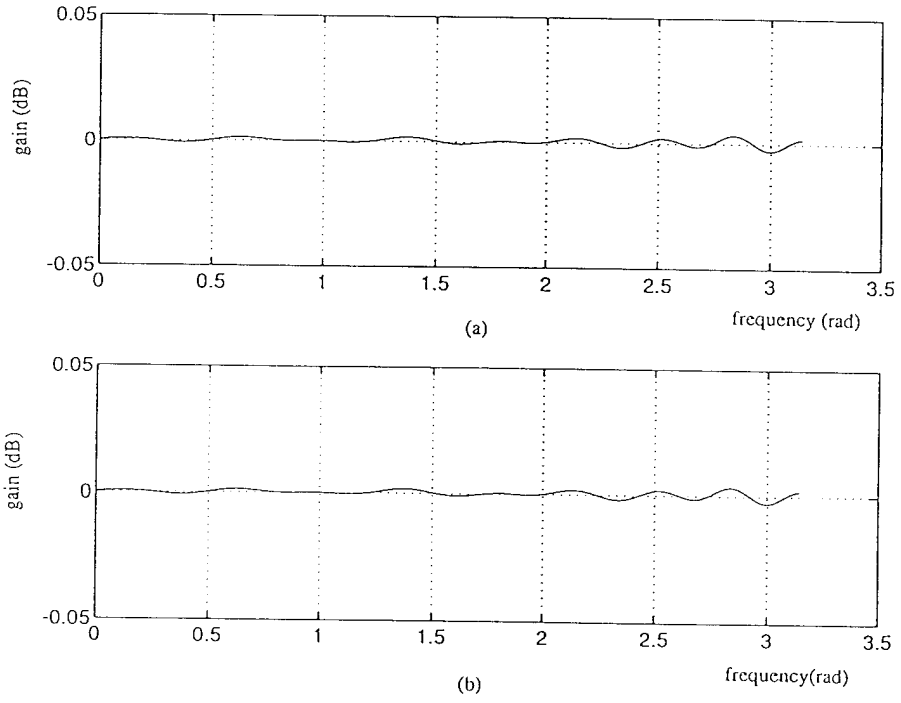


Figure 4: Magnitude response of the 64D Johnston filter polyphase components.
(a) magnitude response of $X_0(z)$ and (b) magnitude response of $X_1(z)$.

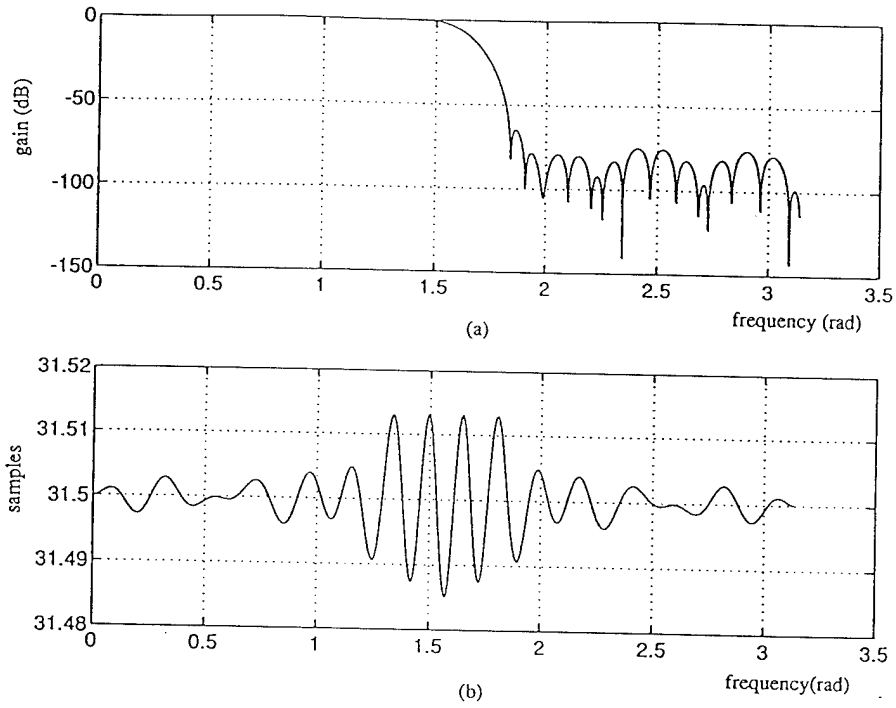


Figure 5: Frequency response of the IIR filter designed in example 1, before optimization.
(a) Magnitude response and (b) group delay response.

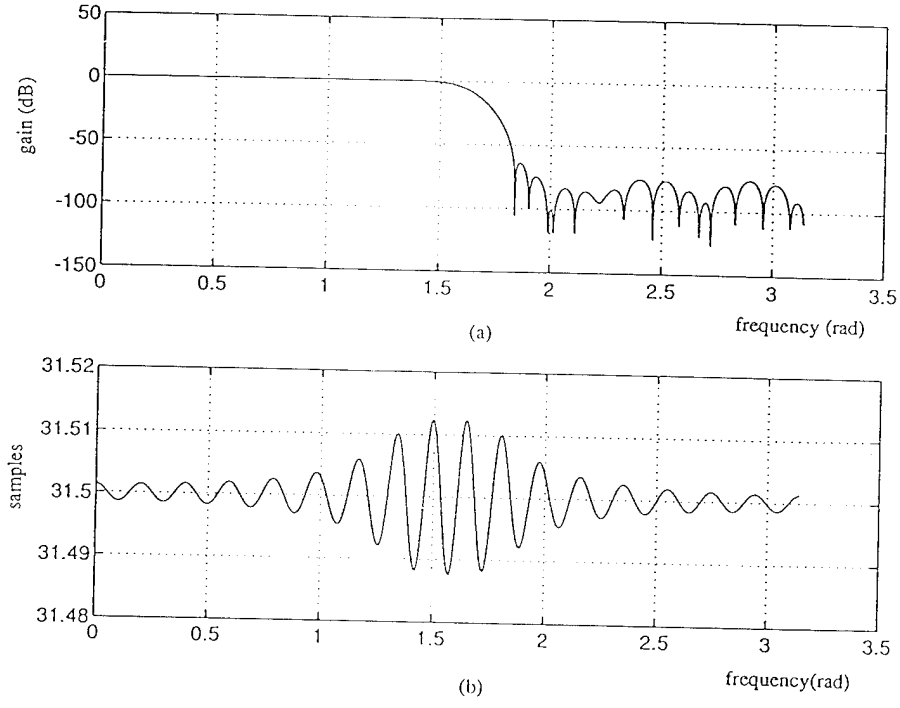


Figure 6: Frequency response of the IIR filter designed in example 1, after optimization
(a) Magnitude response and (b) group delay response.

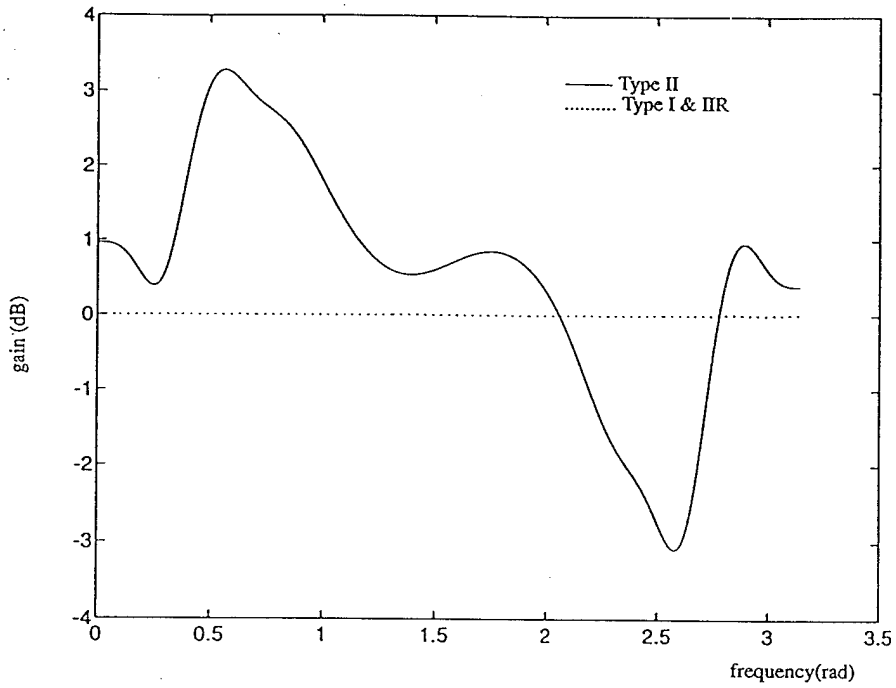


Figure 7: Magnitude response of $T(e^{jw})$ in FWL

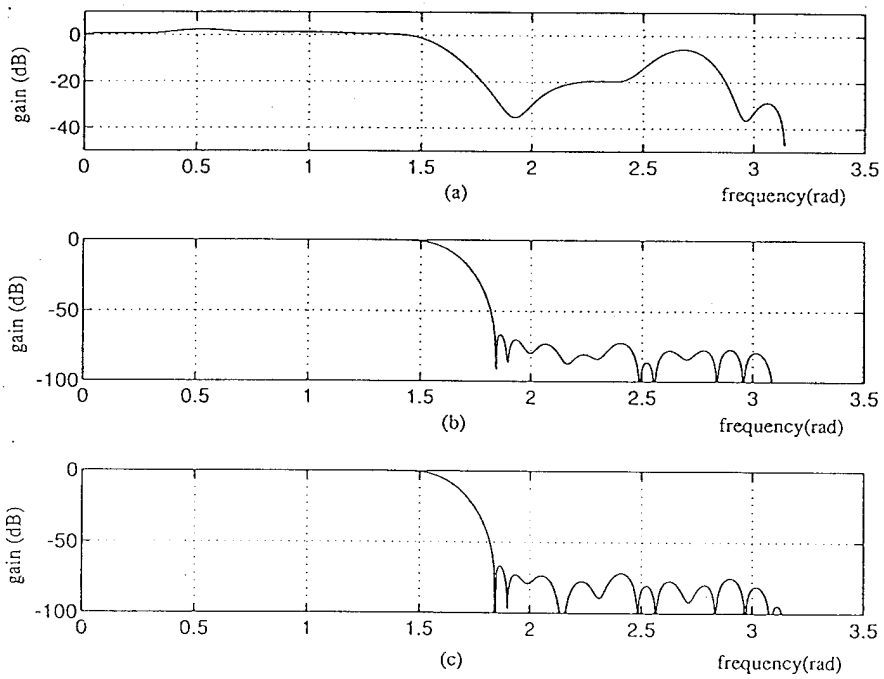


Figure 8: Magnitude response of (a) Type II (b) Type I and (c) IIR filter bank low pass filters when implemented in FWL.

1. INTRODUCTION

The synthesis of digital filters using finite convolution was first proposed by Gold and Jordan [1]. Finite convolution was also used by Voelcker and Hartquist [2] who introduced recursive block processing to exploit the computational advantages of the FFT and other similar block algorithms. Burrus and Parks [3] meanwhile considered the time domain design of recursive digital filters using a matrix formulation of the problem to aid in calculating filter structure.

Block matrix filters (BMFs) provide a state-variable description of block feedback on a matrix implementation of convolution and were first proposed by Burrus [4]. By investigating several block recursion methods, Burrus was also able to demonstrate increased computational efficiency in the processing of digital signals, especially for filters of very high order [5].

The relationship between block implementation of IIR filters as proposed by Burrus and the matrix formulation of IIR filters via direct convolution methods as proposed by Gold and Jordan was investigated by Mitra and Gnansekaran [6] eventually leading to the development of new structures for block implementation of IIR digital filters [7]. Barnes and Shinnaka [8] showed that all irreducible state-space realizations of the matrix filter can be derived through a procedure using a simple realization of the required transfer function. Soon thereafter, Clark, Mitra, and Parker [9] presented a block adaptive filtering procedure using a generalized LMS algorithm for calculating the filter coefficients. In turn, Cioffi [10] applied a deterministic time-domain least-squares criteria within each of the data blocks of the block-adaptive filter to exploit pipelining of the order recursions.

Matrix filters are amenable to analysis and optimization under a variety of criteria. One example is the rank reduction of the least square estimator of a Gaussian process and the resulting improvements in signal-to-noise ratios [11]. Another is the estimation of structured covariance block matrices from stationary time series of multivariate Gaussian processes [12]. Filter matrices can also be constrained so that system limitations are reflected, a topic briefly considered by Ahmed and Rao [13]. The incorporation of matrix structure constraints in the design of MMSE block matrix filters was introduced by Corral and Lindquist [14] [15] and shown to produce computationally efficient forms.

In light of the computational power and efficiency of block matrix filters, it is of general interest to consider those cases when the possible form of the block matrix filter is restricted due to *a priori* constraints. The problem may therefore be summarized as follows:—

> *To find the properties and conditions for the realizability of a block matrix filter given the prescribed constraints, and to design an optimum block matrix*

filter satisfying these constraints.

It is the purpose of this paper to introduce a method for incorporating the practical constraints of system implementation into the design of SISO FIR block matrix filters for time domain digital signal processing. The paper is organized as follows. Section II introduces the basic definitions and nomenclature used in the paper. Section III provides a statement of the MINIMAX ALGORITHM in general and then applies the algorithm to the following matrix structures: time-varying memoryless diagonal, time-invariant periodic circulant, and time-invariant non-periodic Toeplitz. Section IV provides some graphical examples of the algorithm in use with several constraints imposed. Section V provides simulation results for 32×32 matrices and various signal forms in estimation applications using the additive noise model. Section VI considers extensions and limitations of the proposed algorithm. The proofs and properties of the MINIMAX ALGORITHM are relegated to the Appendix.

2. BASIC DEFINITIONS

From a general viewpoint, there are essentially two types of constraints that define the characteristics of a block matrix filter (BMF), namely:

1. Constraints due to the pre-defined rules governing the relationship between the output and the input through the matrix implementation of convolution. These are constraints based on the *operations* of the filter.
2. Constraints due to the design requirements governing the characteristics of the output in terms of the input and the filter matrix. These are constraints based on the *operators* of the filter.

This paper is concerned with item 2. We first begin with some basic definitions.

Definition 2.1. A time domain BMF system is the implementation of the discrete convolution

$$y(n) \stackrel{\text{def}}{=} \sum_{k=0}^{N-1} h(n,k)x(k) \qquad n = 0, 1, \ldots, N-1 \tag{2.1}$$

where $y(n) = \boldsymbol{y} \in \mathbb{R}^{N\times 1}$ is the output vector, $x(k) = \boldsymbol{x} \in \mathbb{R}^{N\times 1}$ is the input vector, and $h(n,k) = h \in \mathbb{R}^{N\times N}$ is the filter matrix.

Eq. (2.1) assumes that we divide the input data stream into data vectors of length N, processing the data vectors in an $N \times N$ system, and then reconstructing the scalar output stream from the processed data vectors. The corresponding single-input single-output (SISO) system is shown in Fig. 2.1 [8].

Definition 2.2. The *matrix structure* constraint is the interrelationship between the elements of the matrix h as

$$h_{i,j} = h_{\rho(i),\sigma(j)} \qquad i,j \in \mathbb{Z} \tag{2.2}$$

where $\rho, \sigma : \mathbb{Z} \mapsto \mathbb{Z}$.

Table 2.1 shows the matrix structures we consider and the corresponding interrelation functions for the indices.

If one of the system requirements is that the filter provide an output within a certain alloted amount of time, or is limited in storage, then we have the following:

Definition 2.3 The *speed-memory* constraint is the *a priori* setting of certain matrix elements to zero as

$$h_{k,l} = 0 \qquad \text{for any } k, l \tag{2.3}$$

For purposes of application, we consider the causal Toeplitz matrix with $h_{i,j} = 0$ for $j > i$ to be a speed-memory constrained matrix.

Let the matrix $h \in \mathbb{R}^{N \times N}$ be non-singular and such that it minimizes some error in

$$\boldsymbol{y} = h\boldsymbol{x} \tag{2.4}$$

Let the matrix $\tilde{h} \in \mathbb{R}^{N \times N}$ be subject to the constraints defined in Eq. (2.2) and (2.3) and its output be given by the relation

$$\tilde{\boldsymbol{y}} = \tilde{h}\boldsymbol{x} = \boldsymbol{y} + \delta\boldsymbol{y} \tag{2.5}$$

The difference between Eqs. (2.4) and (2.5) is the "bias" constraint vector

$$\delta\boldsymbol{y} = \tilde{\boldsymbol{y}} - \boldsymbol{y} = \tilde{h}\boldsymbol{x} - h\boldsymbol{x} = (\tilde{h} - h)\boldsymbol{x} \tag{2.6}$$

Since h is non-singular, we can write

$$\boldsymbol{x} = h^{-1}\boldsymbol{y} \tag{2.7}$$

Therefore, the bias vector is

$$\delta\boldsymbol{y} = (\tilde{h} - h)h^{-1}\boldsymbol{y} = (\tilde{h}h^{-1} - I)\boldsymbol{y} \tag{2.8}$$

where I is the identity matrix. We take the norm of Eq. (2.8) as

$$\|\delta\boldsymbol{y}\| = \|(\tilde{h}h^{-1} - I)\boldsymbol{y}\| \tag{2.9}$$

$$\leq \|\tilde{h}h^{-1} - I\|\|\boldsymbol{y}\| \tag{2.10}$$

where the norm $\|\cdot\|$ satisfies the norm axioms **[16]** and in addition, satisfies the Schwarz inequality **[18]**

$$\|\boldsymbol{v}\boldsymbol{p}\| \leq \|\boldsymbol{v}\|\|\boldsymbol{p}\| \tag{2.11}$$

We are now led to the following:

Definition 2.4 The *relative performance bias* is the measure

$$\frac{\|\delta \boldsymbol{y}\|}{\|\boldsymbol{y}\|} \leq \|\tilde{h}h^{-1} - I\| \tag{2.12}$$

Remark 2.1. The requirement that h be non-singular assumes that in any matrix filter system it is possible to recover the input from the output, i.e., an inverse linear transformation exists.

Definition 2.5 A BMF is *constrained* if any or all of the constraints in **Definition 2.2**, **2.3**, and **2.4** are imposed. A BMF is *realizable* if it meets all the prescribed constraints. A constrained BMF is *optimum* if it is realizable and minimizes the relative performance bias.

For purposes of demonstrating the various constraints, we write $\tilde{h}$ to denote a constrained BMF and $\tilde{\tilde{h}}$ to denote a "further constrained" BMF (i.e., additional constraints are imposed relative to $\tilde{h}$).

Our problem therefore becomes one of determining the realizability of the BMF subject to the prescribed constraints, and to find the optimum based on the minimization of the relative performance bias. Fig. 2.2 shows the flow graph for imposing the various constraints for any given BMF system.

3. STATEMENT OF MINIMAX ALGORITHM

Let $\hat{V}$ be the set of all non-singular matrices and W be the set of constrained matrices. By making the metric $\xi(\tilde{h}h^{-1}, I)$ be induced by the supremum norm we have the relative performance bias of Eq. (2.12), and the equivalent mathematical problem becomes finding $\tilde{h}^{opt} \in W$ among all possible $\tilde{h} \in W$ such that

$$\|\tilde{h}^{opt}h^{-1} - I\|_\infty \leq \|\tilde{h}h^{-1} - I\|_\infty \tag{3.1}$$

for all $h \in \hat{V}$ and where $\|\cdot\|_\infty$ denotes the supremum norm.

From Eq. (3.1) we can construct an equivalent requirement

$$\|\tilde{h}g - I\|_\infty \leq K \tag{3.2}$$

where $g = h^{-1}$ and $K > 0 \in \mathbb{R}$. For the pth row of $\tilde{h}g - I$ we can write the row sums as

$$\sum_{q=0}^{N-1} |\tilde{\boldsymbol{h}}_p \boldsymbol{g}_q - \delta_{p,q}| \leq K \tag{3.3}$$

where $\tilde{\boldsymbol{h}}_p = (\tilde{h}_{p,0}, \tilde{h}_{p,1}, \ldots, \tilde{h}_{p,N-1})$ is the pth row of $\tilde{h}$, $\boldsymbol{g}_q = (g_{0,q}, g_{1,q}, \ldots, g_{N-1,q})^{\mathrm{T}}$ is the qth column of g, and $\delta_{p,q}$ is the Kronecker delta, $\delta_{p,q} = 1$ if $p = q$ and 0 otherwise.

Hence, let

$$\left|\tilde{\boldsymbol{h}}_p \boldsymbol{g}_q - \delta_{p,q}\right| \leq K_{p,q} \tag{3.4}$$

where

$$K_{p,q} = \Delta_{p,q} K \quad \text{with} \quad \sum_{q=0}^{N-1} \Delta_{p,q} = 1 \quad \text{and} \quad \sum_{q=0}^{N-1} K_{p,q} = K \tag{3.5}$$

For any iteration $m > 1$ we can remove the inequalities from Eq. (3.4) and form a set of subconditions as

Condition (Bp) $$-K_{p,p}(m) \leq \tilde{\boldsymbol{h}}_p \boldsymbol{g}_p - 1 \leq +K_{p,p}(m) \tag{3.6}$$

for $p = 0, 1, \ldots, N-1$, and

Condition (Cpq) $$-K_{p,q}(m) \leq \tilde{\boldsymbol{h}}_p \boldsymbol{g}_q \leq +K_{p,q}(m) \quad p \neq q \tag{3.7}$$

for $p, q = 0, 1, \ldots, N-1$, where $K_{p,q}(m)$ is the weighted bound at the mth iteration. We note that $\tilde{\boldsymbol{h}}_p \boldsymbol{g}_q - \delta_{p,q} \leq +K_{p,q}(m)$ and $\tilde{\boldsymbol{h}}_p \boldsymbol{g}_q - \delta_{p,q} \geq -K_{p,q}(m)$ are convex sets for vectors $\tilde{\boldsymbol{h}}_p, \boldsymbol{g}_q$. Indeed, the elements of Eq. (3.4) are closed half-spaces and depend only on the hyperplane established by the equality [17].

Let us denote the convex set established by conditions (Bp) and (Cpq) by Υ. Since Υ is determined by the intersection of a finite set of linear constraints (3.6) and (3.7), the boundary of Υ (if Υ is not empty) will consist of sections of some of the corresponding hyperplanes. Υ will be a region in $\mathbb{R}^D$ ($D \leq N^2$ is the relevant number of coefficients of $\tilde{h}$) and can either be empty, a bounded convex polyhedron, or a convex polyhedron which may be unbounded in some direction in general. If Υ is empty, then no $\tilde{h}$ is realizable; if it is bounded, then the convex polyhedron establishes the region of realizability for $\tilde{h}$ and an optimum exists in the sense of Eq. (3.2); and if Υ is unbounded, then no optimum can be found.

We are therefore led to the following:

REALIZABILITY THEOREM. *If the convex polyhedron established by conditions* (Bp) *and* (Cpq) *for a given K is not empty, then the constrained BMF $\tilde{h}$ is realizable.*

MINIMAX ALGORITHM. *If a filter is realizable, perform the following:*

1. *Select an objective function such as*

$$f_{obj} = \sum_{\rho(i)} \sum_{\sigma(j)} f_{\rho(i),\sigma(j)}(N) \tilde{h}_{\rho(i),\sigma(j)} \tag{3.8}$$

   where $f_{\rho(i),\sigma(j)}(N)$ denotes the number of unconstrained variables $\tilde{h}_{\rho(i),\sigma(j)}$ for each i, j.
2. *Given the bound K, set up conditions* (Bp) *and* (Cpj).
3. *With the above, find the minimum $\tilde{h}^{min}(m)$ and the maximum $\tilde{h}^{max}(m)$ for the mth iteration using the simplex method of linear programming with unconstrained variables* (cf., *Remark* 3.1.).

4. *Calculate the optimum as*

$$\tilde{h}^{opt}(m) = \frac{1}{2}\left(\tilde{h}^{max}(m) + \tilde{h}^{min}(m)\right) \tag{3.9}$$

5. *Calculate the new bound as*

$$K(m+1) = \|\tilde{h}^{opt}(m)g - I\|_{\infty} \tag{3.10}$$

6. *Determine the following:*

   a. *If the bound* $K(m+1) \geq K$, *terminate the process and select* $\tilde{h}^{opt}(m-1)$.

   b. *Else, set* $K = K(m+1)$ *as bound with weights* $\Delta_{p,j} = 1$ *for all* p, j *and go to 2.*

This is a statement of the REALIZABILITY THEOREM and MINIMAX ALGORITHM without proof. The proofs and properties of the REALIZABILITY THEOREM and MINIMAX ALGORITHM are relegated to the Appendix. However, there are several immediate remarks which are directly related to the results and implications of the REALIZABILITY THEOREM and the MINIMAX ALGORITHM.

Remark 3.1. Linear programming problems deal with nonnegative real variables in general. Since $\tilde{h} \in \mathbb{R}^{N\times N}$, we need to use *unconstrained variables* of the form

$$\tilde{h}_{i,j} = \tilde{h}'_{i,j} - \tilde{h}''_{i,j} \quad \tilde{h}'_{i,j}, \tilde{h}''_{i,j} \geq 0 \qquad \text{for all } i,j \tag{3.11}$$

Remark 3.2. Condition (Bp) corresponds to N conditions in general for $p = 0, 1, \ldots, N-1$. Condition (Cpq) corresponds to $N(N-1)$ conditions in general for $p = 0, 1, \ldots, N-1$ and $q = 0, 1, \ldots, N-1$ with $q \neq p$. This corresponds to a total of $N(N-1)+N = N^2$ subconditions for the system of inequalities of Eq. (3.2). If we count each inequality as a separate hyperplane, then we have a total number of $2N^2$ inequality conditions for the two bounding hyperplanes of each inequality enclosing a convex region. Since every admissible domain has only finitely many extreme points, if the domain is defined by r inequalities $x_j \geq 0$, $j = 1, \ldots, r$, and s equations, it can have at most $\binom{r+s}{r} \stackrel{\text{def}}{=} \frac{(r+s)!}{r!s!}$ extreme points [18].

Remark 3.3. It will be shown in Appendix Part I that for the first iterative step $K(1)$ can be set to be arbitrarily large. Ideally, we would like to have $0 \leq K(1) < 1$ imposed *a priori* in Eq. (3.2) because the trivial solution $\tilde{h}^0$ with $\tilde{h}_{i,j} = 0$ for all i, j gives $\|\tilde{h}^0 h^{-1} - I\|_{\infty} = \|I\| = 1$. However, it may be necessary to make the convex polyhedron large enough to hold at least one non-trivial optimum.

Remark 3.4. The requirement of Eq. (3.5) corresponds to a convex combination of $K(m)$ and is obviously closed and convex for any combination of its elements. In order to establish a

non-empty convex polyhedron for conditions (Bp) and (Cpq), it is recommended that $\Delta_{p,j} > 0$. Equal weights are easier to implement in the simplex method [19] and we emphasize their use here.

3.1 Application of Minimax Algorithm to Structural Constraints

The main advantages of the proposed MINIMAX ALGORITHM are, it

1. addresses all the constraints in question, and
2. is based on the well-known theory of linear programming.

These advantages, coupled with the straightforward and systematic procedure for obtaining the optimum BMF once the extreme points of the convex polyhedron are established, make the MINIMAX ALGORITHM highly useful for any constraint of interest. To illustrate the above, let us consider diagonal, circulant, and Toeplitz structural constraints.

Diagonal BMF

For $\tilde{h} \in \mathbb{R}^{N \times N}$ with $h_{i,j} = 0$ for all $i \neq j$, we can rewrite the conditions as

Condition (Bp)
$$1 - K_{p,p}(m) \leq \tilde{h}_{p,p} g_{p,p} \leq 1 + K_{p,p}(m) \tag{3.12}$$
for $p = 0, 1, \ldots, N-1$, and

Condition (Cpq)
$$-K_{p,q}(m) \leq \tilde{h}_{p,p} g_{p,q} \leq +K_{p,q}(m) \tag{3.13}$$
for $p, q = 0, 1, \ldots, N-1$, $p \neq q$. An objective function is

$$f_{obj} = \sum_{i=0}^{N-1} \tilde{h}_{i,i} \tag{3.14}$$

Circulant BMF

For a circulant BMF, $\tilde{h}_{i,j} = \tilde{h}_{mod_N(i+n_0), mod_N(j+n_0)}$ and we have the new conditions

Condition (Bp)
$$1 - K_{p,p}(m) \leq \sum_{i=0}^{N-1} \tilde{h}_{0,i} g_{mod_N(i+p),p} \leq 1 + K_{p,p}(m) \tag{3.15}$$
for $p = 0, 1, \ldots, N-1$, and

Condition (Cpq)
$$-K_{p,q}(m) \leq \sum_{i=0}^{N-1} \tilde{h}_{0,i} g_{mod_N(i+p),q} \leq +K_{p,q}(m) \tag{3.16}$$
for $p, q = 0, 1, \ldots, N-1$, $p \neq q$. An objective function is

$$f_{obj} = \sum_{i=0}^{N-1} N \tilde{h}_{0,i} \tag{3.17}$$

Toeplitz BMF

For a Toeplitz BMF, $\tilde{h}_{i,j} = \tilde{h}_{i+n_0, j+n_0}$ and we have the new conditions

Condition (Bp)
$$1 - K_{p,p}(m) \leq \sum_{i=0}^{p} \tilde{h}_{p-i,0} g_{i,p} + \sum_{i=1}^{p} \tilde{h}_{0,i-p} g_{i,p} \leq 1 + K_{p,p}(m) \tag{3.18}$$

for $p = 0, 1, \ldots, N-1$, and

Condition (Cpq)
$$-K_{p,q}(m) \leq \sum_{i=0}^{N-p-1} \tilde{h}_{0,i} g_{i+p,q} + \sum_{i=1}^{p} \tilde{h}_{i,0} g_{p-i,q} \leq +K_{p,q}(m) \tag{3.19}$$

for $p, q = 0, 1, \ldots, N-1$, $p \neq q$. An objective function is

$$f_{obj} = \sum_{i=0}^{N-1} (N-i)\tilde{h}_{0,i} + \sum_{j=0}^{N-1} (N-j)\tilde{h}_{j,0} \tag{3.20}$$

The conditions (Bp) and (Cpq) and the objective function f_{obj} are used to initialize the simplex tableau for a linear programming solution to the problem using unconstrained variables as stated in *Remark* 3.1. Should different structures be required, the formulations of (Bp) and (Cpq) can be readily modified.

2.2 Application of Minimax Algorithm to Speed-Memory Constraint

Given a block matrix filter must meet a prescribed speed-memory constraint as in **Definition 2.3**, two possible approaches for the incorporation of the constraint in the MINIMAX ALGORITHM are possible, namely: The *A Priori* and *A Posteriori* approaches (cf., Fig. 2.2). The main idea is to set $\tilde{h}_{k,l} = 0$ for select values of k and l. If we incorporate the speed-memory constraint using the *A Posteriori* approach, we can relate the relative performance bias of the further constrained matrix $\tilde{\tilde{h}}$ to the actual output of the original matrix h as follows.

Once the zero elements of $\tilde{\tilde{h}}$ are selected, two paths are possible when implementing the *A Posteriori* approach:

1. Approximate the original matrix h using the formulation

$$\frac{\|\delta\tilde{\tilde{\boldsymbol{y}}}\|_\infty}{\|\boldsymbol{y}\|_\infty} \leq \|\tilde{\tilde{h}}h^{-1} - I\|_\infty \tag{3.21}$$

   where $\delta\tilde{\tilde{\boldsymbol{y}}} = \tilde{\tilde{\boldsymbol{y}}} - \boldsymbol{y}$, $\tilde{\tilde{\boldsymbol{y}}} = \tilde{\tilde{h}}\boldsymbol{x}$.

2. Approximate the constrained matrix $\tilde{h}$ using the bounds

$$\begin{aligned} \frac{\|\delta\tilde{\tilde{\boldsymbol{y}}}\|_\infty}{\|\boldsymbol{y}\|_\infty} &\leq \frac{\|\delta\boldsymbol{y}\|_\infty}{\|\boldsymbol{y}\|_\infty} + \frac{\|\delta\tilde{\boldsymbol{y}}\|_\infty}{\|\boldsymbol{y}\|_\infty} \\ &\leq \|\tilde{h}h^{-1} - I\|_\infty + (1 + \|\tilde{h}h^{-1} - I\|_\infty)\frac{\|\delta\tilde{\boldsymbol{y}}\|_\infty}{\|\tilde{\boldsymbol{y}}\|_\infty} \end{aligned} \tag{3.22}$$

   where $\delta\tilde{\boldsymbol{y}} = \tilde{\tilde{\boldsymbol{y}}} - \tilde{\boldsymbol{y}}$,

$$\frac{\|\delta\tilde{\boldsymbol{y}}\|_\infty}{\|\tilde{\boldsymbol{y}}\|_\infty} \leq \|\tilde{\tilde{h}}\tilde{h}^{-1} - I\|_\infty \tag{3.23}$$

where $\tilde{h}$ is assumed to be non-singular. Here, we have used the fact that

$$\|\tilde{y}\|_\infty \leq (1 + \|\tilde{h}h^{-1} - I\|_\infty)\|y\|_\infty. \tag{3.24}$$

If we want $\dfrac{\|\delta\tilde{\tilde{y}}\|_\infty}{\|y\|_\infty} < 1$, we must consequently assure that

$$\frac{\|\delta\tilde{\tilde{y}}\|_\infty}{\|\tilde{y}\|_\infty} \leq \frac{1 - \|\tilde{h}h^{-1} - I\|_\infty}{1 + \|\tilde{h}h^{-1} - I\|_\infty} \tag{3.25}$$

Remark 3.5. In using procedure 2, we note that the better the original constrained matrix $\tilde{h}$ approximates h, the better the further constrained matrix $\tilde{\tilde{h}}$ approximates h. Eq. (3.25) provides us with a test for the suitability of $\tilde{\tilde{h}}$ in approximating h in general.

4. EXAMPLES USING THE MINIMAX ALGORITHM

° 2×2 **Circulant Example 1**

We consider the approximation of a circulant matrix to itself, namely,

$$\tilde{h} = \begin{pmatrix} \tilde{h}_{0,0} & \tilde{h}_{0,1} \\ \tilde{h}_{0,1} & \tilde{h}_{0,0} \end{pmatrix} \quad h = \begin{pmatrix} 3 & 4 \\ 4 & 3 \end{pmatrix} \quad g = h^{-1} = \begin{pmatrix} -3/7 & 4/7 \\ 4/7 & -3/7 \end{pmatrix} \tag{4.1}$$

We set the weights to be equal, with $\Delta_{p,q} = 1/N = .5$ for all p, q. We also set $K(1) = 10$ so that $K_{p,q}(1) = 5$. From Eqs. (3.15) and (3.16), the set of inequalities becomes

$$\begin{aligned} -4 &\leq -\frac{3}{7}\tilde{h}_{0,0} + \frac{4}{7}\tilde{h}_{0,1} \leq 6 \\ -5 &\leq \frac{4}{7}\tilde{h}_{0,0} - \frac{3}{7}\tilde{h}_{0,1} \leq 5 \\ -5 &\leq \frac{4}{7}\tilde{h}_{0,0} - \frac{3}{7}\tilde{h}_{0,1} \leq 5 \\ -4 &\leq -\frac{3}{7}\tilde{h}_{0,0} + \frac{4}{7}\tilde{h}_{0,1} \leq 6 \end{aligned} \tag{4.2}$$

From Eq. (3.17) the objective function is

$$f_{obj} = 2\tilde{h}_{0,0} + 2\tilde{h}_{0,1} \tag{4.3}$$

with unconstrained variables

$$\tilde{h}_{0,0} = (\tilde{h}'_{0,0} - \tilde{h}''_{0,0}) \quad \tilde{h}_{0,1} = (\tilde{h}'_{0,1} - \tilde{h}''_{0,1}) \quad \tilde{h}'_{0,0}, \tilde{h}''_{0,0}, \tilde{h}'_{0,1}, \tilde{h}''_{0,1} \geq 0 \tag{4.4}$$

Applying the simplex method to the above set of inequalities yields the extrema

$$\text{Min. point : } \tilde{h}_{0,0} = -32, \ \tilde{h}_{0,1} = -31$$

$$\text{Max. point : } \tilde{h}_{0,0} = 38, \ \tilde{h}_{0,1} = 39$$

Eq. (3.9) gives the solution

$$\tilde{h}_{0,0} = \frac{38-32}{2} = 3, \ \tilde{h}_{0,1} = \frac{39-31}{2} = 4 \tag{4.5}$$

This is the original matrix. Figure 4.1 shows the convex polytope for this example.

° 2×2 **Toeplitz Example 2**

° ° *A Priori* approach

We consider the following matrices but let us *a priori* set $\tilde{\tilde{h}}_{0,1} = 0$, requiring

$$\tilde{\tilde{h}} = \begin{pmatrix} \tilde{\tilde{h}}_{0,0} & 0 \\ \tilde{\tilde{h}}_{1,0} & \tilde{\tilde{h}}_{0,0} \end{pmatrix} \quad h = \begin{pmatrix} 5 & -8 \\ 2 & 4 \end{pmatrix} \quad g = h^{-1} = \begin{pmatrix} 1/9 & 2/9 \\ -1/18 & 5/36 \end{pmatrix} \tag{4.6}$$

Initially, let $\Delta_{p,q} = .5$ for all p, q, and let $K(1) = 10$ so that $K_{p,q}(1) = 5$. From Eqs. (3.18) and (3.19), the set of inequalities become

$$\begin{aligned} -4 &\le \frac{1}{9}\tilde{\tilde{h}}_{0,0} \le 6 \\ -5 &\le \frac{2}{9}\tilde{\tilde{h}}_{0,0} \le 5 \\ -5 &\le \frac{1}{9}\tilde{\tilde{h}}_{1,0} - \frac{1}{18}\tilde{\tilde{h}}_{0,0} \le 5 \\ -4 &\le \frac{2}{9}\tilde{\tilde{h}}_{1,0} + \frac{5}{36}\tilde{\tilde{h}}_{0,0} \le 6 \end{aligned} \tag{4.7}$$

Eq. (3.20) gives the objective function

$$f_{obj} = 2\tilde{\tilde{h}}_{0,0} + \tilde{\tilde{h}}_{1,0} \tag{4.8}$$

Applying the MINIMAX ALGORITHM gives the following matrix in the first iteration

$$\tilde{\tilde{h}}(1) = \begin{pmatrix} 0 & 0 \\ 4.5 & 0 \end{pmatrix} \quad \text{with} \quad \tilde{\tilde{h}}(1)h^{-1} - I = \begin{pmatrix} -1 & 0 \\ .5 & 0 \end{pmatrix} \tag{4.9}$$

so that $\|\tilde{\tilde{h}}h^{-1} - I\|_\infty = 1$. We cannot improve this answer because the selection of $\tilde{h}_{0,1} = 0$ has made the trivial solution $\tilde{h} = 0$ the optimum solution.

° °*A Posteriori* approach

In the *A Posteriori* approach, we first calculate the full approximating matrix $\tilde{h}$ and then calculate the further constrained matrix $\tilde{\tilde{h}}$ based either on the constrained matrix $\tilde{h}$ or the unconstrained matrix h. Applying the MINIMAX ALGORITHM we find that the optimum constrained matrix is

$$\tilde{h} = \begin{pmatrix} 5 & -8 \\ 1.375 & 5 \end{pmatrix} \tag{4.10}$$

One possible heuristic for selecting the proper value $\tilde{\tilde{h}}_{k,l}$ to set to zero is to select the value closest to zero in magnitude. Consequently, we would make $\tilde{\tilde{h}}_{1,0} = 0$. We also note that if we wanted to approximate to $\tilde{h}$ (or h), we must satisfy Eq. (3.25), that is,

$$\frac{\|\delta\tilde{\boldsymbol{y}}\|}{\|\tilde{\boldsymbol{y}}\|} \leq \frac{1-.125}{1+.125} = .7778$$

Let us first attempt to approximate to $\tilde{h}$, requiring

$$\tilde{\tilde{h}} = \begin{pmatrix} \tilde{\tilde{h}}_{0,0} & \tilde{\tilde{h}}_{0,1} \\ 0 & \tilde{\tilde{h}}_{0,0} \end{pmatrix} \quad \tilde{h} = \begin{pmatrix} 5 & -8 \\ 1.375 & 4 \end{pmatrix} \quad \tilde{g} = \tilde{h}^{-1} = \begin{pmatrix} .1389 & .2222 \\ -.0382 & .1389 \end{pmatrix} \tag{4.11}$$

Let $\Delta_{p,q} = .5$ for all p, q, and let $K(1) = 10$ so that $K_{p,q}(1) = 5$. From Eqs. (3.18) and (3.19), the set of inequalities now becomes

$$\begin{aligned} -4 &\leq .1389\tilde{\tilde{h}}_{0,0} - .0382\tilde{\tilde{h}}_{0,1} \leq 6 \\ -5 &\leq .2222\tilde{\tilde{h}}_{0,0} + .1389\tilde{\tilde{h}}_{0,1} \leq 5 \\ -5 &\leq -.0382\tilde{\tilde{h}}_{0,0} \leq 5 \\ -4 &\leq .1389\tilde{\tilde{h}}_{0,0} \leq 6 \end{aligned} \tag{4.12}$$

Eq. (3.20) gives the objective function

$$f_{obj} = 2\tilde{\tilde{h}}_{0,0} + \tilde{\tilde{h}}_{0,1} \tag{4.13}$$

Applying the MINIMAX ALGORITHM gives the following matrix in the first iteration

$$\tilde{\tilde{h}}(1) = \begin{pmatrix} 5 & -8 \\ 0 & 5 \end{pmatrix} \quad \text{with} \quad \tilde{\tilde{h}}(1)\tilde{h}^{-1} - I = \begin{pmatrix} 0 & 0 \\ -.1910 & -.3055 \end{pmatrix} \tag{4.14}$$

so that $\|\tilde{\tilde{h}}(1)\tilde{h}^{-1} - I\| \leq .4965$. We can improve the answer by additional iterations, obtaining the optimum as

$$\tilde{\tilde{h}}^{opt} = \begin{pmatrix} 5.934 & -9.778 \\ 0 & 5.934 \end{pmatrix} \tag{4.15}$$

with the relative performance bias $\|\tilde{\tilde{h}}^{opt} h^{-1} - I\|_\infty \leq .5055$.

Alternatively, we can approximate to the original unconstrained h, requiring

$$\tilde{\tilde{h}} = \begin{pmatrix} \tilde{\tilde{h}}_{0,0} & \tilde{\tilde{h}}_{0,1} \\ 0 & \tilde{\tilde{h}}_{0,0} \end{pmatrix} \quad \tilde{h} = \begin{pmatrix} 5 & -8 \\ 2 & 4 \end{pmatrix} \quad \tilde{g} = \tilde{h}^{-1} = \begin{pmatrix} 1/9 & 2/9 \\ -1/18 & 5/36 \end{pmatrix} \tag{4.16}$$

Applying the MINIMAX ALGORITHM gives the following optimum Toeplitz matrix

$$\tilde{\tilde{h}}^{opt} = \begin{pmatrix} 6.0466 & -9.67456 \\ 0 & 6.0466 \end{pmatrix} \tag{4.17}$$

so that $\|\tilde{h}^{opt} h^{-1} - I\|_\infty \leq .4961$. This is very close to the result of Eq. (4.15).

5. SIMULATION RESULTS USING MINIMAX ALGORITHM

In communications systems data is often transmitted in some digital modulation format such as BPSK, QPSK, QAM, etc., with a sinusoidal carrier. Determining the presence of the proper format via estimation, detection, or correlation techniques is critical to minimize Bit Error Rates (BER). In addition to typical sinusoidal waveforms, it is of general interest to investigate the performance of any proposed method with signals exhibiting sharp transitions. Consequently, we provide simulations for both sinusoidal and non-sinusoidal waveforms in the difficult application of signal estimation in noisy environment.

For purposes of simulation, we consider estimation application with non-singular matrices h such that $h\boldsymbol{x} = \boldsymbol{d}$, that is, the matrix being approximated provides the desired output from the input. For the constrained approximating matrix $\tilde{h}$ we considered the following forms:

a. Time-varying memoryless diagonal.

b. Time-invariant periodic circulant.

c. Time-invariant non-periodic Toeplitz.

d. Time-invariant non-periodic cuasal Toeplitz.

The simulations were implemented in FORTRAN on a VAX 4000-600. The signals simulated had $N = 32$ samples with period $T = 1$. The additive white noise model was used with a signal-to-noise ratio of 10dB. One group of signals has two periodic signals—the sine and cosine waveform with period=.19635 for one full cycle in the 32-sample window. The other group has two non-periodic signals—the ramp with an increment=.03125 for each sample and the exponential with a damping factor=.125. Plots of the matrices h, actual input and desired output for the corresponding matrices are given in Figures 5.1–5.4.

In addition to the termination requirements of the MINIMAX ALGORITHM, we further imposed the following scheme: If the error was being reduced but by a value less than a threshold τ, that is, if

$$\|\tilde{h}^{opt}(m)h^{-1} - I\|_\infty - \|\tilde{h}^{opt}(m-1)h^{-1} - I\|_\infty < \tau \tag{5.1}$$

the procedure was terminated and the optimum solution for the previous iteration was used.

For our simulations, an original value of $\tau = .05$ was selected. However, it was often necessary to modify the above procedure to prematurely terminate the iterative process allowing for non-trivial solutions, even if these violate the relative performance bias bound of unity. This is due to the fact that for large matrices, the optimality condition is hard to obtain (see Section VI). The IMSL routine DLPRS **[20]** was used for the MINIMAX ALGORITHM.

Fig. 5.5 shows the plots of the optimum constrained diagonal matrix, the actual output, and the relative performance bias bound for each signal type. For the diagonal matrix, the output contains more noise than the input, so performance is inferior.

Fig. 5.6 shows the plots of the optimum constrained circulant matrix, the actual output, and the relative performance bias bound for each signal type. The circulant matrices perform well, especially for periodic signals as expected. The actual relative performance bias shows that the MINIMAX ALGORITHM is able to extract the essential features of the input from the matrix being approximated.

Fig. 5.7 shows the plots of the optimum constrained Toepltiz matrix, the actual output, and the relative performance bias bound for each signal type. The Toeplitz matrix filters perform well in general, but the remarkable feature of this simulation is that for periodic signals, the best Toeplitz structure is the circulant structure. This is due to the fact that the MINIMAX ALGORITHM is approximating the same type of general matrix for the periodic signals, and that the circulant structure is a special case of the Toeplitz structure.

Fig. 5.8 shows the plots of the optimum constrained causal Toeplitz matrix, the actual output, and the relative performance bias bound for each signal type. The causal Toeplitz matrix can be viewed as a Toeplitz matrix with a speed-memory constraint imposed such that $h_{i,j} = 0$ for $j > i$. Although the performance is inferior to the circulant and Toeplitz filters, it is better than the diagonal filters while still only storing N elements.

6. EXTENSIONS AND LIMITATIONS OF MINIMAX ALGORTIHM

If the MINIMAX ALGORITHM is adjusted for minimizing the column sum the result is a minimization of the corresponding 1-norm of the output bias. Consequently, the REALIZABILITY THEOREM and MINIMAX ALGORITHM address both norms through a simple change in formulation.

A cursory analysis of the MINIMAX ALGORITHM would reveal that the main problem is addressed and solved through an iterative procedure that is computationally intensive: There are N^2 subconditions in general, and there are as many as $2N$ variables for the matrices we are considering. However, we can note that $N^2 - f_{\rho(i),\sigma(j)}(N)$ conditions are redundant. As a result of applying the simplex method, these redundant conditions are never considered, thereby reducing the actual number of computations. For the speed-memory constraint, additional variables are eliminated, further reducing the computational load.

If we increase N, however, for the majority of practical situations will yield $K(m) > 1$ for any m. The reasons are outlined below:

1. The matrix $\tilde{h}$ is not well-suited to approximate h. This is especially true for speed-

memory constrained matrices.

2. In order to satisfy $\|\tilde{h}h^{-1} - I\|_\infty < 1$, each subcondition must satisfy

$$\sum_{q=0}^{N-1} |\tilde{\boldsymbol{h}}_p(m)\boldsymbol{g}_q - \delta_{p,q}| < 1 \tag{6.1}$$

for each row p. As N increases, it becomes more difficult to satisfy Eq. (6.1).

Given the fact that for large N the result may be the trivial solution $\tilde{h}^0$, it may not be necessary to terminate the process. In the Appendix it is shown that as long as certain conditions are satisfied, the MINIMAX ALGORITHM reduces the bound K. While an optimum solution with bound $K > 1$ may not be desirable in absolute terms, it may be "tolerable" for the application at hand.

The supremum norm is the maximum of the row sum of the relative performance bias. It is possible that one row, say row k, gives the condition $K > 1$ but all other rows $i = 1, 2, \ldots$, $i \neq k$ satisfy $K < 1$. The resulting output will therefore have its maximum error at $y(k)$. By additional post-processing it may be possible to reduce the error at $y(k)$ and thereby still satisfy the realizability conditions. This is also applicable to the reduction of the maximum of the column sums of the relative performance bias for the 1-norm problem.

In the simulations of Section V we calculated the optimum $\tilde{h}$ even under the condition that $K > 1$. This is to demonstrate the method while still keeping in mind that the trivial solution may be necessary in order to insure $K = 1$. If $K \gg 1$, then the trivial solution $\tilde{h}^0$ may be the only solution.

Although the MINIMAX ALGORITHM finds the solution via the relative performance bias bound, it can be extended to the more traditional methods of minimizing the error of the output vector from the desired output vector without any loss of generality. Consider the minimization of the error

$$\epsilon = \tilde{\boldsymbol{y}} - \boldsymbol{d} = \tilde{h}\boldsymbol{x} - \boldsymbol{d} \tag{6.2}$$

where $\tilde{\boldsymbol{y}}$ is the actual output vector and $\boldsymbol{d}$ is the desired output vector. There exists at least one $h \in \mathbb{R}^{N\times N}$ such that $h\boldsymbol{x} = \boldsymbol{d}$. Furthermore, if h is non-singular, then $\boldsymbol{x} = h^{-1}\boldsymbol{d}$. Therefore, the error can be written as

$$\epsilon = \tilde{h}\boldsymbol{x} - h\boldsymbol{x} = (\tilde{h} - h)\boldsymbol{x} = (\tilde{h} - h)h^{-1}\boldsymbol{d} \tag{6.3}$$

Taking the supremum norm of both sides and simplifying, we get the main result

$$\|\epsilon\|_\infty \leq \|\tilde{h}h^{-1} - I\|_\infty \|\boldsymbol{d}\|_\infty \tag{6.4}$$

We can minimize $\|\epsilon\|_\infty$ by minimizing the relative performance bias.

7. CONCLUSION

We have introduced the REALIZABILITY THEOREM for testing the realizability of a BMF subject to the matrix structure, speed-memory, and relative performance bias constraint. The MINIMAX ALGORITHM uses the REALIZABILITY THEOREM in an iterative procedure to find the optimum BMF subject to the constraints.

The MINIMAX ALGORITHM is based on the simplex method of linear programming applied to a system of linear constraints obtained from the supremum norm of the relative performance bias. The optimum at each iterative step is the midway point between the minimum and the maximum points of the convex polytope established by the linear constraints. For each iterative step, a new bound is calculated from the optimum and used for the next iteration. It has been shown that this iterative technique always reduces the bounds as long as the convex polytope established by the linear constraints is not empty.

The MINIMAX ALGORITHM was shown to be extendable to a variety of error parameters under a generalized approach. In addition, it was shown that the application of the above constraints were straightforward without any loss of generality. Examples and simulation results show that the MINIMAX ALGORITHM can be applied to the design of optimum block matrix filters subject to the prescribed system constraints.

APPENDIX: PROOFS AND PROPERTIES

The Appendix is concerned with the proofs and properties of the REALIZABILITY THEOREM and MINIMAX ALGORITHM proposed in Section III. The Appendix is organized as follows. Part I provides the proofs and properties of the REALIZABILITY THEOREM. Part II provides a detailed analysis of the constituent parts of the MINIMAX ALGORITHM while Part III provides the convergence proof of the algorithm.

Part I. REALIZABILITY THEOREM

The REALIZABILITY THEOREM establishes the existence of an optimum, or equivalently, a region of realizability. In order for an optimum to exist, the convex polyhedron Υ established by Conditions (Bp) and (Cpq) must be bounded in all directions. We must consequently have

THEOREM 1.1. *For any finite $K \geq 0$, conditions* (Bp) *and* (Cpq) *establish a convex polyhedron that is either empty or bounded in all directions.*

Proof. The condition for empty is trivial, so let us prove the bounded case. The matrix $g = h^{-1}$ used to establish conditions (Bp) and (Cpq) is the inverse of the matrix h that is being approximated. Therefore, the columns of g are linearly independent and none of the

intersecting pair of bounding hyperplanes established by $\tilde{\boldsymbol{h}}_p \boldsymbol{g}_q - \delta_{p,q}$ are parallel to each other for any p or q. Therefore, they all intersect and bound the resulting convex polyhedron. □

The domain of realizability is a bounded convex polyhedron (*convex polytope*) established by the prescribed constraints. We will now show that if an optimum exists for a nonempty convex polytope constructed from a given set of weights and an initial bound, the optimum exists for any bound $K' > K$.

THEOREM 1.2. *Given a set of weights $\Delta_{p,q}$ such that $K_{p,q} = \Delta_{p,q} K$ and the optimum exists for this value of K, then the optimum exists for any $K' > K$.*

Proof. For any p and q, the bounds are

$$-K_{p,q} \leq \tilde{\mathbf{h}}_p \boldsymbol{g}_q - \delta_{p,q} \leq +K_{p,q} \quad \text{or} \quad 0 \leq \tilde{\mathbf{h}}_p \boldsymbol{g}_q - \delta_{p,q} + K_{p,q} \leq 2K_{p,q} \tag{1.1}$$

Now let $K' = \alpha K$ where $1 \leq \alpha < \infty$. Also, $K'_{p,q} = \Delta_{p,q} K'$. Substituting into Eq. (1.1) we get

$$0 \leq \tilde{\mathbf{h}}_p \boldsymbol{g}_q - \delta_{p,q} + K''_{p,q} \leq 2K''_{p,q} \tag{1.2a}$$

$$0 \leq \tilde{\mathbf{h}}_p \boldsymbol{g}_q - \delta_{p,q} + \alpha K_{p,q} \leq 2\alpha K_{p,q} \tag{1.2b}$$

Dividing Eq. (1.2b) through by α we obtain

$$0 \leq \frac{1}{\alpha}(\tilde{\mathbf{h}}_p \boldsymbol{g}_q - \delta_{p,q}) + K_{p,q} \leq 2K_{p,q} \tag{1.3}$$

From Eq. (1.1) we get the main result

$$0 \leq \frac{1}{\alpha} K_{p,q} + K_{p,q} \leq 2K_{p,q} \tag{1.4}$$

or

$$0 \leq \frac{\alpha + 1}{\alpha} \leq 2 \tag{1.5}$$

which is true for $1 \leq \alpha < \infty$. □

Remark 1.1. THEOREM 1.2 basically states that our ability to find an optimum is not sensitive to the initial bound K. (We will show this in more detail in the development of the actual method in Section IV.) K can be set arbitrarily large in order to assure we enclose at least one non-trivial optimum (cf., *Remark* 3.3 in Section III).

A useful result is the following:

THEOREM 1.3. *If $\|\tilde{h}^{opt} g - I\|_\infty = K'$ for some $K > K'$, then the optimum exists for any set of weights $\Delta_{p,q}$ such that $K_{p,q} = \Delta_{p,q} K > K'$ for all p, q and the given K.*

Proof. By THEOREM 1.2, we can select the initial bound K arbitrarily large. Since Υ is non-empty for any $K > \|\tilde{h}^{opt}g - I\|_\infty$, we can make K such that $K \to K'$ and Υ is still not empty. Moreover, we can make $K_{p,q} \to K'$ for any p, q. The set of weights satisfying this relation, i.e., $\Delta_{p,q} > K'/K$, assure Υ is not empty and consequently the optimum exists. □

Remark 1.2. The optimum does not exist for any set of weights in general, only those specified by THEOREM 1.3, namely, the weights that guarantee that the convex polytope Υ is not empty. Since any of the above set of weights can be used, it is obvious that uniform weights $\Delta_{p,q}$ (i.e., $\Delta_{p,q} = \Delta_{r,s} > K'/K$ for all p, q, r, s) can be used with equal effectiveness.

The optimum block matrix filter therefore exists and can be determined from the convex polytope Υ. However, the optimum is not unique in general because of the supremum norm's *max* operator. Additional constraints need to be imposed in order to find a global optimum. We will now show that the MINIMAX ALGORITHM finds the optimum through the iterative procedure described in Section III. We first begin with a brief analysis of the constituent parts of the MINIMAX ALGORITHM.

Part II. THE OBJECTIVE FUNCTION

In linear programming problems, the objective function f_{obj} is either minimized or maximized based on the linear constraints given. For the MINIMAX ALGORITHM, the objective function must have two basic properties:

1. The objective function must be a linear function of all the variables of $\tilde{h}$ in question. This guarantees that all the extreme points of the convex polytope Υ will be considered.
2. The objective function's formulation must result in a hyperplane that is not parallel to any of the bounding hyperplanes that constitute the convex polytope Υ.

The objective function is "swept" through the convex polytope Υ to obtain either the minimum or maximum extreme point.

The MINIMAX ALGORITHM addresses the realizability of optimum block matrix filters with real elements. Consequently, the extreme points of Υ obtained by f_{obj} can be positive or negative (this is the reason we need unconstrained variables). An objective function of the form

$$f_{obj} = \beta\tilde{\boldsymbol{h}} = \beta_{0,0}\tilde{h}_{0,0} + \beta_{0,1}\tilde{h}_{0,1} + \cdots + \beta_{1,0}\tilde{h}_{1,0} + \cdots \qquad \beta_{k,l} > 0 \tag{2.1}$$

where $\tilde{\boldsymbol{h}} = (\tilde{h}_{0,0}, \tilde{h}_{0,1}, \ldots, \tilde{h}_{1,0}, \ldots)^{\mathrm{T}}$ with $\tilde{h}_{k,l} = (\tilde{h}'_{k,l} - \tilde{h}''_{k,l})$ denotes the vector of relevant variables of the matrix $\tilde{h}$ and $\beta = (\beta_{0,0}, \beta_{0,1}, \ldots, \beta_{1,0}, \ldots)$ is the coefficient vector, assures that if there are negative (positive) extreme vertices in Υ, these will result in a minimum (maximum) solution for f_{obj}.

Although Eq. (2.1) satisfies property 1, we must select the proper $\beta_{k,l} > 0$ to satisfy property 2. It is easy to see that if we make the set $\{\beta\}$ countable by letting $\beta_{k,l} \in \mathbb{N}$, then it is a countable subset of $\mathbb{R}$ and hence is of measure zero [21]. This assures property 2, that is, for the set of all possible hyperplanes $\tilde{\boldsymbol{h}}_p \boldsymbol{g}_q - \delta_{p,q}$ with $g, \tilde{h} \in \mathbb{R}^{N \times N}$, $\beta\tilde{\boldsymbol{h}}$ is of measure zero.

The authors selected an objective function in which the coefficients $\beta_{k,l}$ were set to be equal to the number of elements for the variable $\tilde{h}_{k,l}$, that is, in proper form,

$$\beta_{\rho(i),\sigma(j)} = f_{\rho(i),\sigma(j)}(N) \qquad \text{all } i,j \tag{2.2}$$

This corresponds to an objective function in which the variables are weighted by the number of times they appear in the matrix $\tilde{h}$ (e.g., Eqs. (3.14), (3.17), and (3.20) in Section III).

Since we are working with unconstrained variables, it is easy to show that

$$\frac{\partial f_{obj}}{\partial \tilde{h}'_{\rho(i),\sigma(j)}} = -\frac{\partial f_{obj}}{\partial \tilde{h}''_{\rho(i),\sigma(j)}} \tag{2.3}$$

that is, the derivatives are equal but in opposite directions. With the weighting coefficients $\beta_{k,l}$ we force the largest positive and negative elements of the given convex polytope via the prescribed objective function.

The weights are selected such that if we form the ratio f_{obj}/N^2 and sum the resultant coefficients for each matrix structure, we get 1. For the diagonal matrices, each coefficient is a uniform weight as $\frac{1}{N}\tilde{h}_{i,i}$, $i = 0, 1, \ldots, N-1$ and since there are N variables, the result is 1. For circulant matrices, each coefficient is a uniform weight as $\frac{1}{N}\tilde{h}_{0,i}$, $i = 0, 1, \ldots, N-1$, and since there are again N variables, the sum is also 1.

Toeplitz matrices have the coefficients weighted as $\frac{N-1}{N^2}\tilde{h}_{0,i}$ and $\frac{N-1}{N^2}\tilde{h}_{i,0}$ for $i = 0, 1, \ldots, N-1$. The sum of the coefficients also result in 1:

$$sum = \frac{1}{N} + \sum_{i=1}^{N-1} 2\left(\frac{N-i}{N^2}\right) \tag{2.4}$$

where $\frac{1}{N}$ is from the element $\tilde{h}_{0,0}$ and $2\left(\frac{N-i}{N^2}\right)$ is a result of summing the coefficients for each element $\tilde{h}_{0,i}$ and $\tilde{h}_{i,0}$. From Eq. (2.4) we have

$$\begin{aligned} sum &= \frac{1}{N} + \sum_{i=1}^{N-1} 2\left(\frac{N-i}{N^2}\right) \\ &= \frac{1}{N} + \frac{2}{N^2}\sum_{i=1}^{N-1}(N-i) = \frac{1}{N} + \frac{2}{N^2}\left[\sum_{i=1}^{N-1} N - \sum_{i=1}^{N-1} i\right] \\ &= \frac{1}{N} + \frac{2}{N^2}\left[N(N-1) - \frac{N(N-1)}{2}\right] = \frac{1}{N} + \frac{2}{N^2}\left[\frac{2N^2 - 2N - N^2 + N}{2}\right] \\ &= \frac{1}{N} + \frac{N^2 - N}{N^2} = 1 \end{aligned}$$

Part III. THE MINIMAX ALGORITHM

The procedure for determining the realizability of a constrained block matrix filter requires constructing a convex polyhedron Υ bounded by $K(1) > 0$. By THEOREM 1.1, Υ is a convex polytope. By THEOREM 1.2, $K(1)$ can be chosen arbitrarily large. Since we have an intersection of hyperplanes we note that each initial subcondition can be represented by the relation

$$\tilde{h}_p^{max}(1)g_q - \delta_{p,q} = +K_{p,q}(1) - \gamma_{p,q}^{max}(1) \tag{3.1a}$$

$$\tilde{h}_p^{min}(1)g_q - \delta_{p,q} = -K_{p,q}(1) + \gamma_{p,q}^{min}(1) \tag{3.1b}$$

where

$$0 \leq \gamma_{p,q}^{min}(1), \gamma_{p,q}^{max}(1) \leq 2K_{p,q}(1) \tag{3.2}$$

constitute the offsets resulting from the actual intersection of the hyperplanes. If we let the initial optimum be calculated as a convex combination of the extreme points as

$$\tilde{h}^{opt}(2) = \frac{1}{c}\tilde{h}^{max}(1) + \frac{c-1}{c}\tilde{h}^{min}(1) \quad c > 1 \tag{3.3}$$

then we can show the following

LEMMA 3.1. *If* $0 \leq \gamma_{p,q}^{min}(1), \gamma_{p,q}^{max}(1) \leq K_{p,j}(1)$ *then*

$$\tilde{h}_p^{opt}(2)g_q - \delta_{p,q} \leq \tilde{h}_p^{max}(1)g_q - \delta_{p,q} \quad \text{and} \quad \tilde{h}_p^{opt}(2)g_q - \delta_{p,q} \geq \tilde{h}_p^{min}(1)g_q - \delta_{p,q} \tag{3.4}$$

and if $K_{p,q}(1) \leq \gamma_{p,q}^{min}(1), \gamma_{p,q}^{max}(1) \leq 2K_{p,q}(1)$ *then*

$$\tilde{h}_p^{opt}(2)g_q - \delta_{p,q} \geq \tilde{h}_p^{max}(1)g_q - \delta_{p,q} \quad \text{and} \quad \tilde{h}_p^{opt}(2)g_q - \delta_{p,q} \leq \tilde{h}_p^{min}(1)g_q - \delta_{p,q} \tag{3.5}$$

Remark 3.1. LEMMA 3.1 states that if we take the convex combination of the two extreme vertices of Υ, we are never outside the convex polytope, and furthermore, we improve the answer. This result is trivial at first, but its proof details the iterative procedure and leads to establishing a value for c.

Proof. The optimum for any subcondition is

$$\begin{aligned}\tilde{h}_p^{opt}(2)g_q - \delta_{p,q} &= \left[\frac{1}{c}\tilde{h}_p^{max}(1) + \frac{c-1}{c}\tilde{h}_p^{min}(1)\right]g_q - \delta_{p,q} \\ &= \frac{1}{c}\tilde{h}_p^{max}(1)g_q + \frac{c-1}{c}\tilde{h}_p^{min}(1)g_q - \delta_{p,q}\end{aligned} \tag{3.6}$$

in general. Substituting Eqs. (3.1) into Eq. (3.6) and rearranging we get

$$\tilde{h}_p^{opt}(2)g_q - \delta_{p,q} = \frac{-c+2}{c}K_{p,q}(1) - \frac{1}{c}\gamma_{p,q}^{max}(1) + \frac{c-1}{c}\gamma_{p,q}^{min}(1) \tag{3.7}$$

If $0 \leq \gamma_{p,q}^{min}(1), \gamma_{p,q}^{max}(1) \leq K_{p,q}(1)$ we want to show that

$$\tilde{\boldsymbol{h}}_p^{opt}(2)\boldsymbol{g}_q - \delta_{p,q} \leq \tilde{\boldsymbol{h}}_p^{max}(1)\boldsymbol{g}_q - \delta_{p,q}$$

or, from Eq. (3.7) and Eq. (3.1a),

$$\frac{-c+2}{c}K_{p,q}(1) - \frac{1}{c}\gamma_{p,q}^{max}(1) + \frac{c-1}{c}\gamma_{p,q}^{min}(1) \leq K_{p,q}(1) - \gamma_{p,q}^{max}(1) \tag{3.8}$$

Assume the opposite is true, then multiplying Eq. (3.8) by c and rearranging we get

$$(c-1)\gamma_{p,q}^{max}(1) + (c-1)\gamma_{p,q}^{min}(1) \geq (2c-2)K_{p,q}(1) \tag{3.9a}$$

$$(c-1)\left[\gamma_{p,q}^{max}(1) + \gamma_{p,q}^{min}(1)\right] \geq (c-1)2K_{p,q}(1) \tag{3.9b}$$

or

$$\gamma_{p,q}^{max}(1) + \gamma_{p,q}^{min}(1) \geq 2K_{p,q}(1) \tag{3.9c}$$

But this is a contradiction since $0 \leq \gamma_{p,q}^{min}(1), \gamma_{p,q}^{max}(1) \leq K_{p,q}(1)$, so we have proven Eq. (3.8). The same argument can be followed for the other conditions by substituting the appropriate assumptions and inequalities in Eqs. (3.9). □

LEMMA 3.1 can be extended to any iteration m for the nonempty convex polytope Υ. From Eq. (3.8), we see that if we set $c = 2$, we eliminate $K_{p,q}(m)$ from each iteration subcondition, that is, for any iteration m, and letting

$$\tilde{\boldsymbol{h}}^{opt}(m+1) = \frac{1}{2}\left(\tilde{\boldsymbol{h}}^{max}(m) + \tilde{\boldsymbol{h}}^{min}(m)\right) \tag{3.10}$$

then

$$\tilde{\boldsymbol{h}}_p^{opt}(m+1)\boldsymbol{g}_q - \delta_{p,q} = \frac{1}{2}\left(\gamma_{p,q}^{min}(m) - \gamma_{p,q}^{max}(m)\right) \qquad \text{any } p,q \tag{3.11}$$

If any $\gamma_{p,q}^{min}(m) = \gamma_{p,q}^{max}(m)$, $\tilde{\boldsymbol{h}}_p^{opt}(m+1)\boldsymbol{g}_q - \delta_{p,q}$ is identically zero, which corresponds to collapsing the convex polyhedron for that particular p and q to a single hyperplane. At the next iteration, the convex polytope is *a fortiori* empty.

We further note that there are other cases for which $\tilde{\boldsymbol{h}}^{opt}(m)\boldsymbol{g}_q - \delta_{p,q}$ is reduced to a single hyperplane. Consider the case when $\gamma_{k,l}^{min}(m) = 0$ and $\gamma_{k,l}^{max}(m) = 2K_{p,q}(m)$ for some index k,l and iteration m, then Eqs. (3.1) become

$$\tilde{\boldsymbol{h}}_k^{max}(m)\boldsymbol{g}_l - \delta_{k,l} = +K_{k,l}(m) - 2K_{k,l}(m) = -K_{k,l}(m) \tag{3.12a}$$

$$\tilde{\boldsymbol{h}}_k^{min}(m)\boldsymbol{g}_l - \delta_{k,l} = -K_{k,l}(m) + 0 = -K_{k,l}(m) \tag{3.12b}$$

which also corresponds to a single hyperplane. Furthermore, $\tilde{\boldsymbol{h}}_k^{max} = d\tilde{\boldsymbol{h}}_k^{min}$ if $\tilde{\boldsymbol{h}}_k^{min}$ is normal to $\boldsymbol{g}_l$, or $d = 1$ otherwise. It can also be shown that we have a single hyperplane if $\gamma_{k,l}^{min}(m) =$

$2K_{k,l}(m)$ and $\gamma_{k,l}^{max}(m) = 0$. Therefore, for any nonempty convex polytope Υ, we must necessarily have

$$0 < \gamma_{k,l}^{min}(m), \gamma_{k,l}^{max}(m) < 2K_{p,q}(m) \tag{3.13}$$

LEMMA 3.2. *If the convex polytope Υ is not empty for the given m, then*

$$K_{p,q}(m) < K_{p,q}(m-1) \qquad m \geq 2 \tag{3.14}$$

where $K_{p,q}(m) = \max_{p,q} |\tilde{h}^{opt}(m)g - I|$.

Proof. The optimum at iteration $m+1$ yields the result

$$\tilde{\mathbf{h}}_p^{opt}(m+1)\boldsymbol{g}_q - \delta_{p,q} = \frac{1}{2}(\gamma_{p,q}^{min}(m) - \gamma_{p,q}^{max}(m)) \qquad \text{any } p,q \tag{3.10}$$

The next bound is selected as

$$K_{p,q}(m+1) = \max_{p,q} |\tilde{h}^{opt}(m+1)g - I|$$

and is some positive value, necessarily of the form

$$K_{p,q}(m+1) = \frac{1}{2}|\gamma_{p_m,q_m}^{min}(m) - \gamma_{p_m,q_m}^{max}(m)| \quad \text{all } p,q \tag{3.15}$$

where the subscripts p_m, q_m denote the indices for which $\max_{p_m,q_m} |\tilde{h}^{opt}(m+1)g - I| \geq \max_{p,q} |\tilde{h}^{opt}(m+1)g - I|$ for all p,q at the mth iteration.

Reinitializing the bounds we have

$$\tilde{\mathbf{h}}_p^{max}(m+1)\boldsymbol{g}_q - \delta_{p,q} = +K_{p,q}(m+1) - \gamma_{p,q}^{max}(m+1) \tag{3.16a}$$

$$\tilde{\mathbf{h}}_p^{min}(m+1)\boldsymbol{g}_q - \delta_{p,q} = -K_{p,q}(m+1) + \gamma_{p,q}^{min}(m+1) \tag{3.16b}$$

and $0 < \gamma_{p,q}^{min}(m+1), \gamma_{p,q}^{max}(m+1) < 2K_{p,q}(m+1)$ and $K_{p,q}(m+1)$ is given by Eq. (3.11). In the next iterative step we have

$$\begin{aligned}\tilde{\mathbf{h}}_p^{opt}(m+2)\boldsymbol{g}_q - \delta_{p,q} &= \frac{1}{2}(\gamma_{p,q}^{min}(m+1) - \gamma_{p,q}^{max}(m+1)) \\ &< \frac{1}{2}\left|\gamma_{p_m,q_m}^{min}(m) - \gamma_{p_m,q_m}^{max}(m)\right|\end{aligned}$$

and consequently

$$K_{p,q}(m+2) = \frac{1}{2}\left|\gamma_{p_{m+1},q_{m+1}}^{min}(m+1) - \gamma_{p_{m+1},q_{m+1}}^{max}(m+1)\right| < K_{p,q}(m+1) \quad \text{all } p,q \tag{3.17}$$

The argument can be extended for $K_{p,q}(m+i)$, $i = 1, 2, \ldots$, as long as the convex polytope is not empty.

THEOREM 3.3. *For any indices p, q and iteration m, if $\tilde{\boldsymbol{h}}_p^{opt}(m)\boldsymbol{g}_q - \delta_{p,q}$ does not reduce any subcondition to a single hyperplane, then*

$$\|\tilde{h}^{opt}(m)g - I\|_\infty < \|\tilde{h}^{opt}(m-1)g - I\|_\infty \tag{3.18}$$

Proof. Follows from LEMMA 3.2. Since Υ is non-empty, we are guaranteed that we reduce the bounds for each subcondition. Since $\|\tilde{h}^{opt}(m-1)g - I\|_\infty = \max_p \sum_q |\tilde{\boldsymbol{h}}_p^{opt}(m-1)\boldsymbol{g}_q - \delta_{p,q}|$ and $|\tilde{\boldsymbol{h}}_p^{opt}(m)\boldsymbol{g}_q - \delta_{p,q}| < |\tilde{\boldsymbol{h}}_p^{opt}(m-1)\boldsymbol{g}_q - \delta_{p,q}|$ for all p, q, then it is also true for the sums and Eq. (3.18) is obviously true. □

$K_{p,q}(m)$ is a monotonically nonincreasing function. The iterative procedure continues until we collapse one or more of the bounding convex polyhedra to single hyperplanes or we force the convex polytope to be empty. When the bounding convex polyhedra are reduced to single hyperplanes, we proceed with one more iteration step. This *can* result in an improved solution although the convex polytope is now empty and consequently, no more iterations are possible as there are no feasible solutions for the simplex method.

It is at this juncture where the MINIMAX ALGORITHM departs from the usual procedure for showing the convergence of a sequence of iterations since the strongest result we can obtain is THEOREM 3.3, namely, that $K_{p,q}(m)$ is a monotonically nonincreasing function for nonempty Υ. Because in the last iteration step the answer can be improved although the convex polytope is empty, it is not possible to show that for every case

$$\lim_{m\to\infty} \max_p \sum_q K_{p,q}(m) = K' \tag{3.19}$$

This ambiguity is not inconsistent with the procedure as at each iterative step the same process takes place. It is the purpose of the MINIMAX ALGORITHM to collapse the bounding convex polyhedra to single hyperplanes.

REFERENCES

[1] B. Gold and K. L. Jordan, "A note on digital filter synthesis," Proc. IEEE (Lett.), vol. 56, pp. 1717-1718, Oct. 1968.

[2] H. B. Voelcker and E. E. Hartquist, "Digital filtering via block recursion," IEEE Trans. on Audio Electroacoust., vol. AU-18, No. 2, pp. 169–176, June 1970.

[3] C. S. Burrus and T. W. Parks, "Time domain design of recursive digital filters," IEEE Trans. on Audio Electroacoust., vol. AU-18, No. 2, pp. 137–141, June 1970.

[4] C. S. Burrus, "Block implementation of digital filters," IEEE Trans. on Circuit Theory, vol. CT-18, No. 6, pp. 697–701, Nov. 1971.

[5] C. S. Burrus, "Block realization of digital filters," IEEE Trans. on Audio Electroacoust., vol. AU-20, No. 4, pp. 230–235, Oct. 1972.

[6] R. Gnansekaran and S. K. Mitra, "A note on block implementation of IIR digital filters," Proc. IEEE (Lett.), vol. 65, pp. 1063–1064, July 1977.

[7] S. K. Mitra and R. Gnansekaran, "Block implementation of recursive digital filters—new structures and properties," IEEE Trans. Circ. Syst., vol. CAS-25, No. 4, pp. 200–207, April 1978.

[8] C. W. Barnes and S. Shinnaka, "Block-shift invariance and block implementation of discrete-time filters," IEEE Trans. Circ. Syst., vol. CAS-27, No. 8, pp. 667–672, Aug. 1980.

[9] G. A. Clark, S. K. Mitra, and S. R. Parker, "Block implementation of adaptive digital filters," IEEE Trans. on Acoust. Speech and Sig. Proc., vol. ASSP-29, No. 3, pp. 744–752, June 1981.

[10] J. M. Cioffi, "The block-processing FTF adaptive algorithm," IEEE Trans. on Acoust. Speech and Sig. Proc., vol. ASSP-34, No. 1, pp. 77–90, Feb. 1986.

[11] L. L. Scharf, "Topics in Statistical Signal Processing," §6.9 in J. L. Lacoume, T. S. Durrani, and R. Stora (Eds.), *Signal Processing, Volume 1*, New York: North-Holland, 1987.

[12] J. P. Burg, D. G. Luenberger, and D. L. Wenger, "Estimation of structured covariance matrices," Proc. IEEE, vol. 70, No. 9, pp. 963–974, Sep. 1982.

[13] N. Ahmed and K. R. Rao, *Orthogonal Transforms for Digital Signal Processing*, New York: Springer-Verlag, 1975.

[14] C. A. Corral and C. S. Lindquist, "Design of constrained minimum mean-square error digital matrix filters," 26th Asilomar Conf. on Signals, Systems and Computers, pp. 596–600, Oct. 1992.

[15] C. A. Corral and C. S. Lindquist, "Analysis and design of structurally constrained minimum mean-square error block matrix filters," submitted to IEEE Trans. on Sig. Proc. (Corresp.), December 1993.

[16] A. S. Householder, *The Theory of Matrices in Numerical Analysis*, New York: Blaisdell Pub. Co., 1964.

[17] R. T. Rockafellar, *Convex Analysis*, Princeton, NJ: Princeton University Press, 1970.

[18] I. N. Bronshtein and K. A. Semendyayev, *Handbook of Mathematics*, New York: Van Nostrand Reinhold Co., 1985.

[19] C. A. Corral, *Analysis and Design of Optimum Block Matrix Filters with Prescribed System Constraints*, unpublished Ph.D. dissertation, University of Miami, Coral Gables, Florida, Dec. 1993.

[20] IMSL MATH/LIBRARY Manual, §8.3, "Linearly Constrained Optimization", pp. 888–891, based on R. J. Hanson and J. A. Wisniewski, *A revised simplex code for LP problems using orthogonal decomposition—A User's Guide*, Sandia Labs, Technical Report SAND78-2322.

[21] R. R. Goldberg, *Methods of Real Analysis*, New York: Blaisdell Publishing Co., 1964.

FIGURES

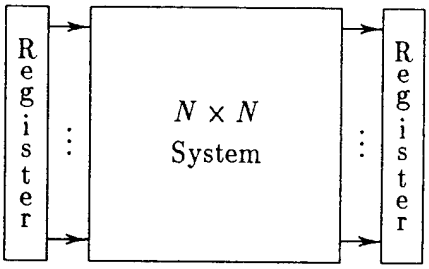


Figure 2.1 $N \times N$ **SISO System.**

Table 2.1 Matrix Structure and Interrelation Functions

Matrix Structure	Interrelation Function
Diagonal	$\rho(i) = i,\ \sigma(j) = j$
Circulant	$\rho(i) = mod_N(i + n_0),\ \sigma(j) = mod_N(j + n_0)$
Toeplitz	$\rho(i) = i + n_0,\ \sigma(j) = j + n_0$

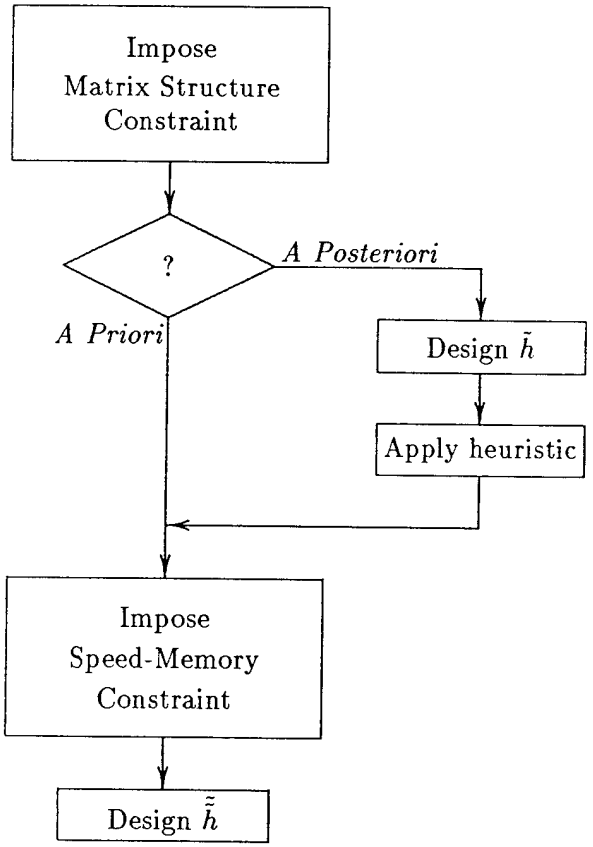


Figure 2.2 Flow graph for imposing constraints on system.

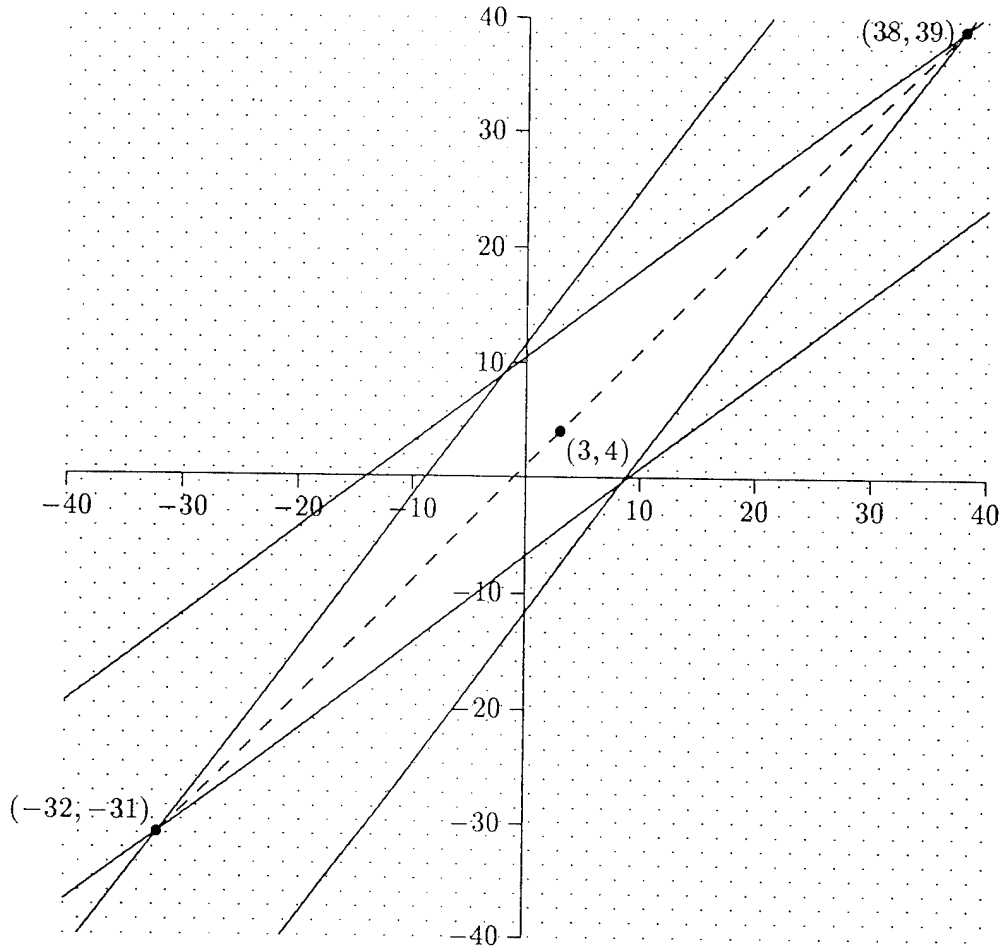
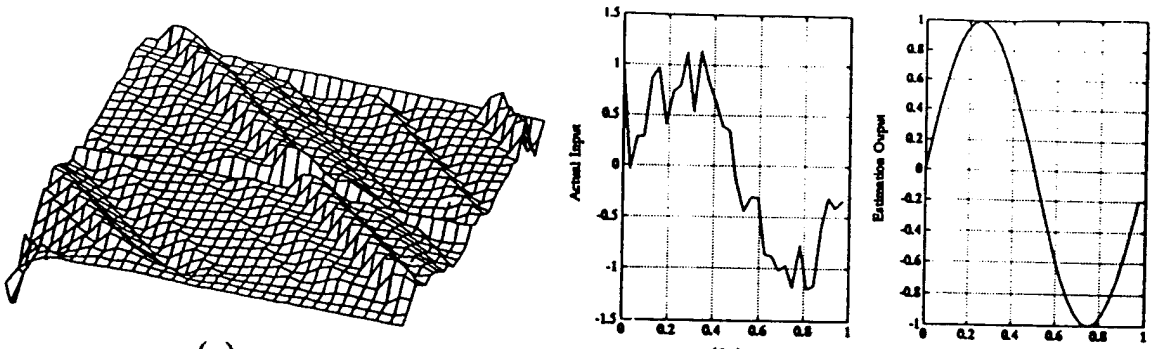
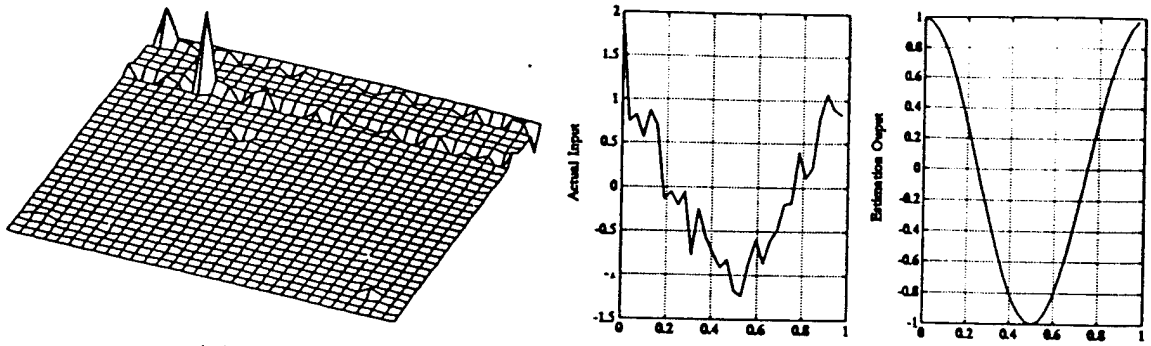


Figure 4.1. Convex polytope for 2×2 Circulant Example 1. Unshaded area is the region of feasible solutions. The minimum, maximum, and midpoint solution are shown.



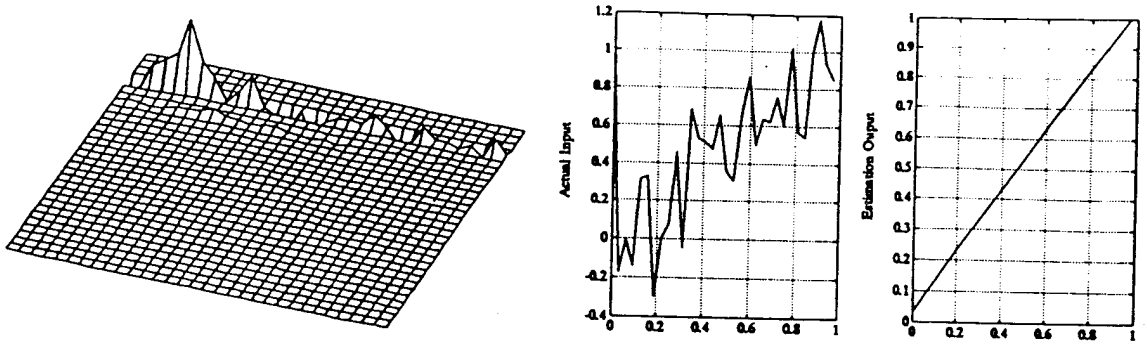
(a) (b) (c)

Figure 5.1 (a) Estimation matrix for sine wave, (b) input, SNR= 10dB, (c) output.



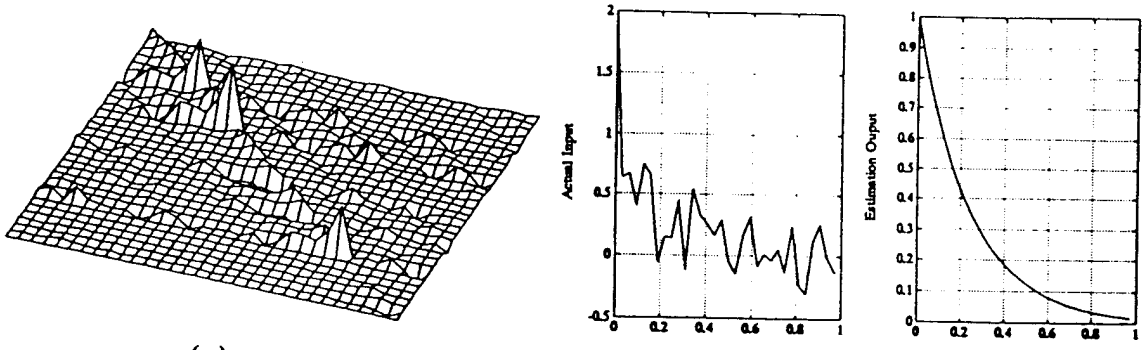
(a) (b) (c)

Figure 5.2 (a) Estimation matrix for cosine wave, (b) input, SNR= 10dB, (c) output.



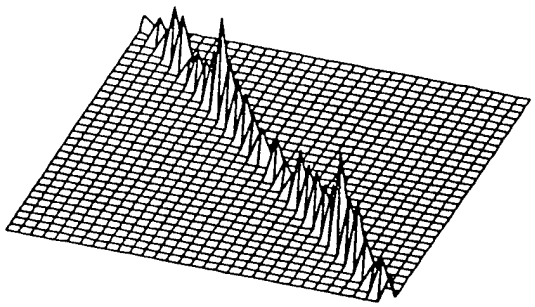
(a) (b) (c)

Figure 5.3 (a) Estimation matrix for ramp signal, (b) input, SNR= 10dB, (c) output.

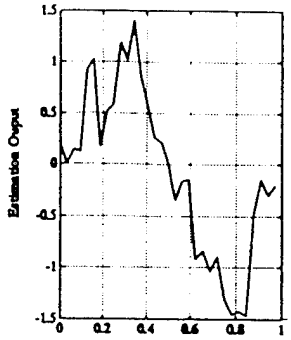


(a) (b) (c)

Figure 5.4 (a) Estimation matrix for exp signal, (b) input, SNR= 10dB, (c) output.



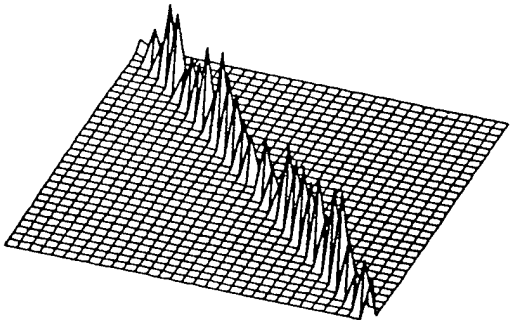
(a)



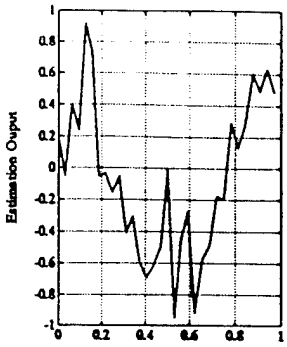
(b)

$$\frac{\|\delta y\|_\infty}{\|y\|_\infty} \leq .516$$

(c)



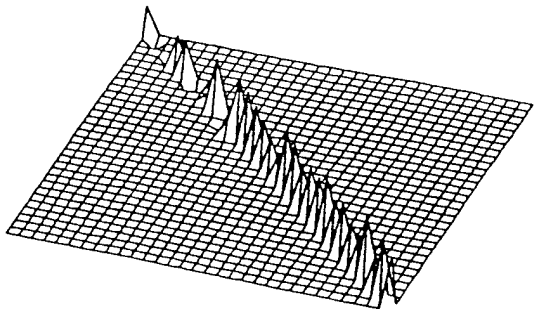
(a)



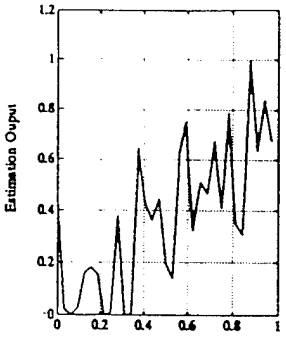
(b)

$$\frac{\|\delta y\|_\infty}{\|y\|_\infty} \leq 1.08$$

(c)



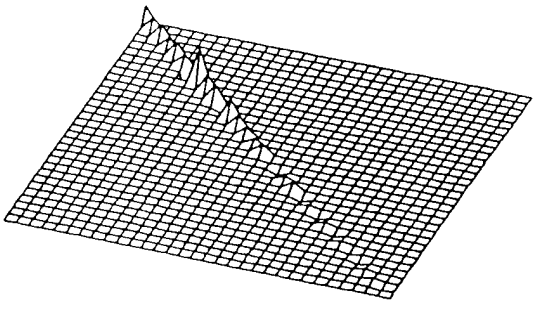
(a)



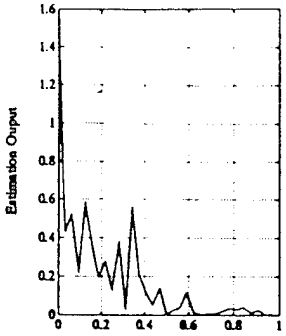
(b)

$$\frac{\|\delta y\|_\infty}{\|y\|_\infty} \leq .568$$

(c)



(a)

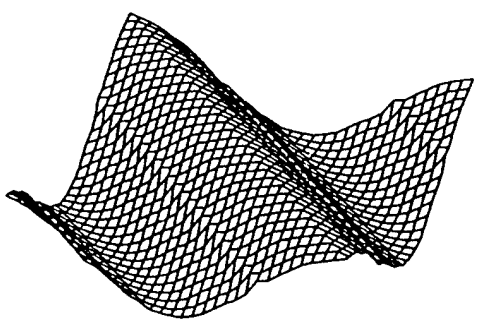


(b)

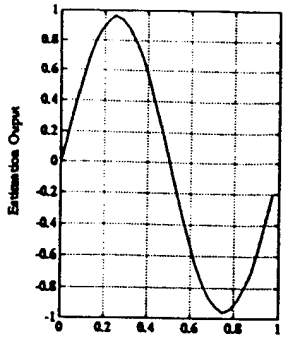
$$\frac{\|\delta y\|_\infty}{\|y\|_\infty} = .375$$

(c)

Figure 5.5 Optimum constrained diagonal filter $\tilde{h}$ for sin, cos, t, **and** exp **signal, respectively: (a) Estimation matrix, (b) estimation output and (c) relative performance bias bounds, for each signal type.**



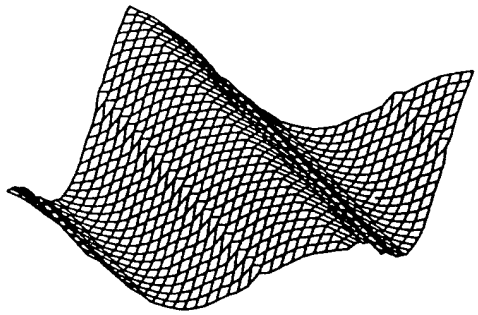
(a)



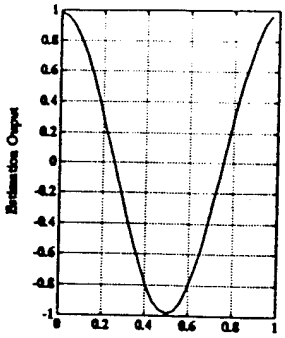
(b)

$$\frac{\|\delta y\|_\infty}{\|y\|_\infty} \leq .054$$

(c)



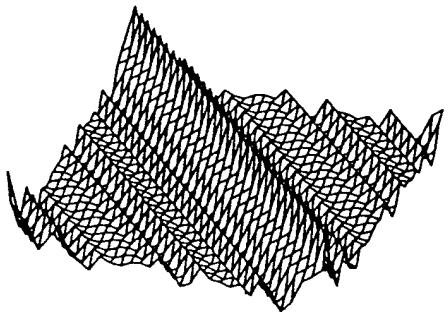
(a)



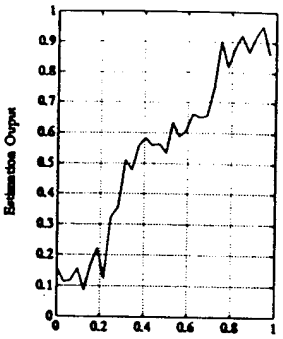
(b)

$$\frac{\|\delta y\|_\infty}{\|y\|_\infty} \leq .033$$

(c)



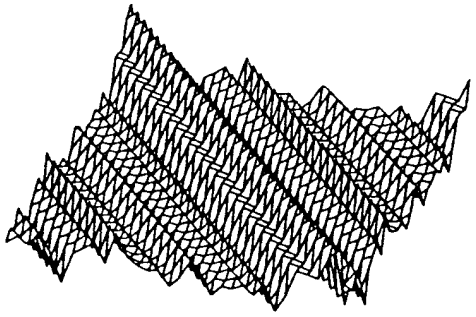
(a)



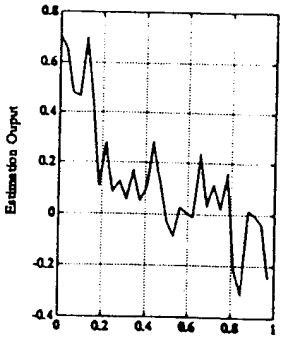
(b)

$$\frac{\|\delta y\|_\infty}{\|y\|_\infty} \leq .175$$

(c)



(a)



(b)

$$\frac{\|\delta y\|_\infty}{\|y\|_\infty} \leq .492$$

(c)

Figure 5.6 Optimum constrained circulant filter $\tilde{h}$ for sin, cos, t, and exp signal, respectively: (a) Estimation matrix, (b) estimation output and (c) relative performance bias bounds, for each signal type.

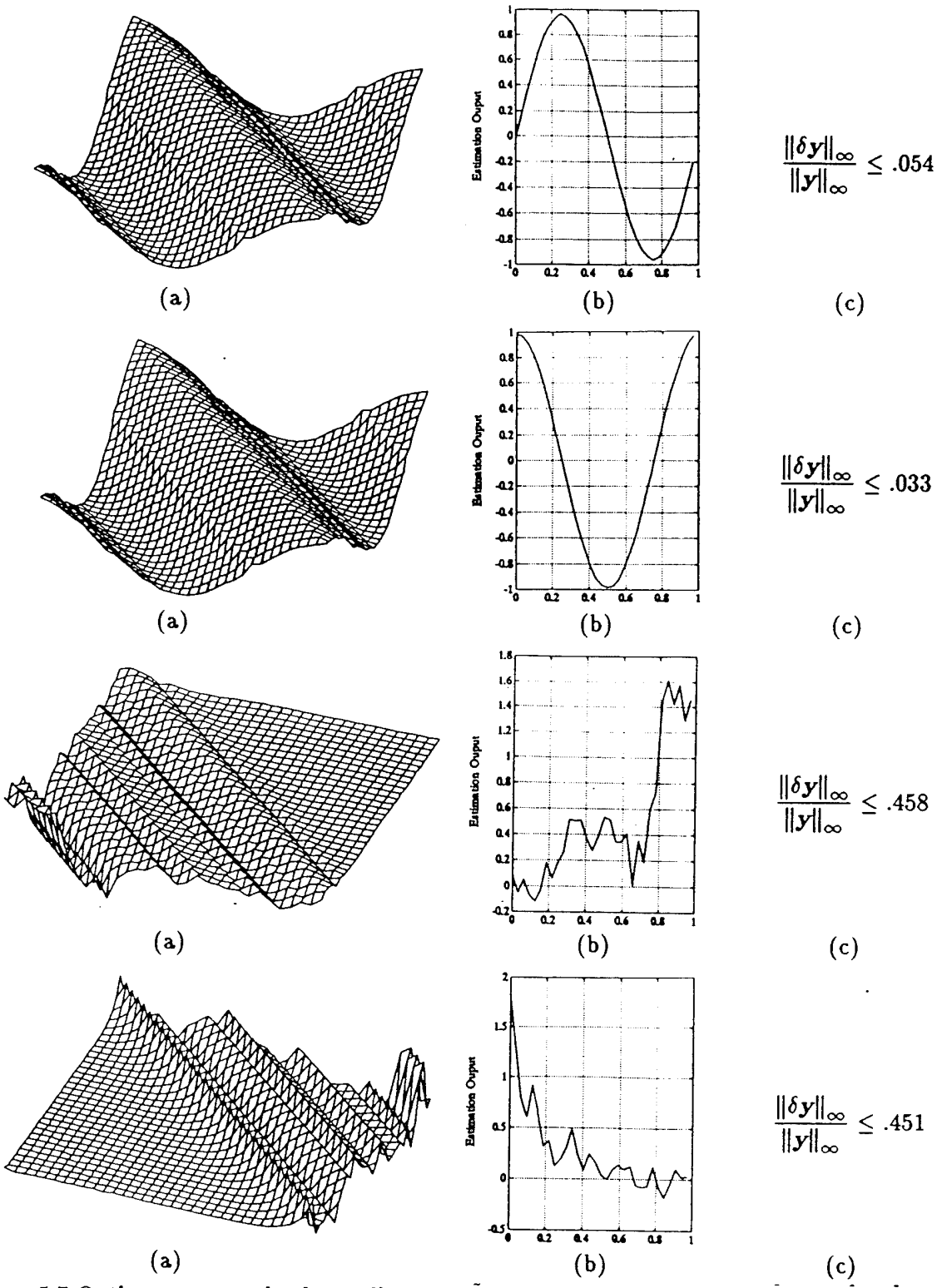
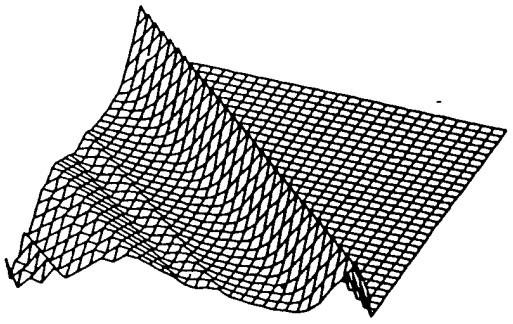
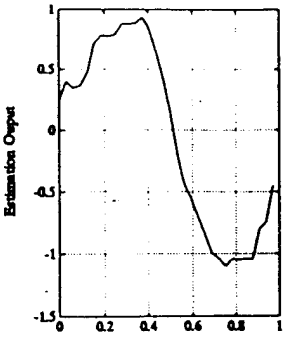


Figure 5.7 Optimum constrained Toeplitz filter $\tilde{h}$ for sin, cos, t, and exp **signal, respectively: (a) Estimation matrix, (b) estimation output and (c) relative performance bias bounds, for each signal type.**



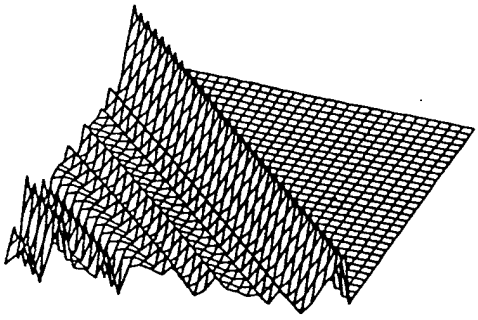
(a)



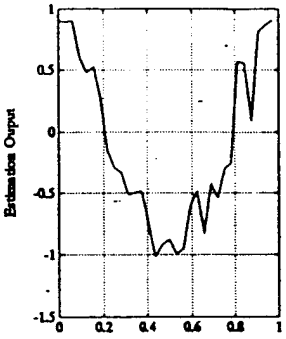
(b)

$$\frac{\|\delta y\|_\infty}{\|y\|_\infty} \leq .3283$$

(c)



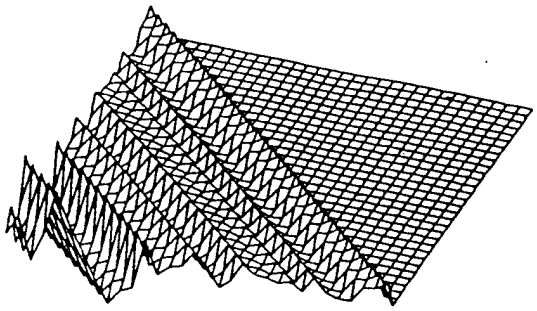
(a)



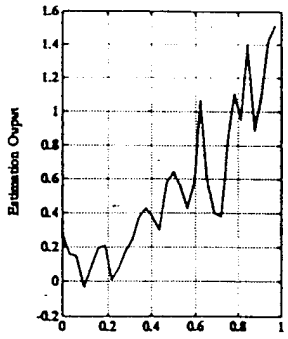
(b)

$$\frac{\|\delta y\|_\infty}{\|y\|_\infty} \leq .6064$$

(c)



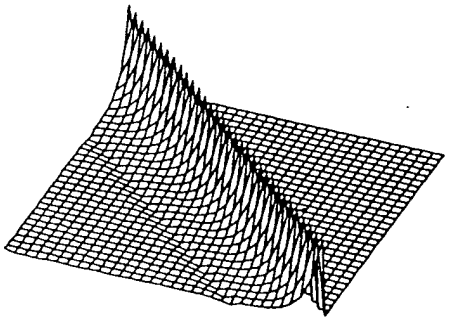
(a)



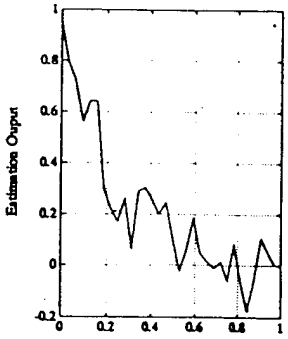
(b)

$$\frac{\|\delta y\|_\infty}{\|y\|_\infty} \leq .3476$$

(c)



(a)



(b)

$$\frac{\|\delta y\|_\infty}{\|y\|_\infty} \leq .3181$$

(c)

Figure 5.8 Optimum constrained causal Toeplitz filter $\tilde{h}$ for sin, cos, t, **and** exp **signal, respectively: (a) Estimation matrix, (b) estimation output and (c) relative performance bias bounds, for each signal type.**